VISIT…

# Advanced Information and Knowledge Processing

**Series Editors**
Professor Lakhmi Jain
Lakhmi.jain@unisa.edu.au
Professor Xindong Wu
xwu@cs.uvm.edu

For other titles published in this series, go to
www.springer.com/series/4738

Dan A. Simovici · Chabane Djeraba

# Mathematical Tools for Data Mining

## Set Theory, Partial Orders, Combinatorics

Springer

Dan A. Simovici, MS, MS, PhD
University of Massachusetts, Boston
USA

Chabane Djeraba, BSc, MSc, PhD
University of Sciences and Technologies
of Lille (USTL)
France

AI&KP ISSN 1610-3947
ISBN: 978-1-84800-200-5 e-ISBN: 978-1-84800-201-2
DOI: 10.1007/978-1-84800-201-2

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2008932365

©Springer-Verlag London Limited 2008
Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licenses issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.
The use of registered names, trademarks, etc., in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.
The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springer.com

# Preface

This volume was born from the experience of the authors as researchers and educators, which suggests that many students of data mining are handicapped in their research by the lack of a formal, systematic education in its mathematics.

The data mining literature contains many excellent titles that address the needs of users with a variety of interests ranging from decision making to pattern investigation in biological data. However, these books do not deal with the mathematical tools that are currently needed by data mining researchers and doctoral students. We felt it timely to produce a book that integrates the mathematics of data mining with its applications. We emphasize that this book is about mathematical tools for data mining and *not* about data mining itself; despite this, a substantial amount of applications of mathematical concepts in data mining are presented. The book is intended as a reference for the working data miner.

In our opinion, three areas of mathematics are vital for data mining: *set theory*, including partially ordered sets and combinatorics; *linear algebra*, with its many applications in principal component analysis and neural networks; and *probability theory*, which plays a foundational role in statistics, machine learning and data mining.

This volume is dedicated to the study of set-theoretical foundations of data mining. Two further volumes are contemplated that will cover linear algebra and probability theory.

The first part of this book, dedicated to set theory, begins with a study of functions and relations. Applications of these fundamental concepts to such issues as equivalences and partitions are discussed. Also, we prepare the ground for the following volumes by discussing indicator functions, fields and $\sigma$-fields, and other concepts.

In this part, we have also included a précis of universal and linear algebra that covers the needs of subsequent chapters. This part concludes with a chapter on graphs and hypergraphs.

The second part is centered around partially ordered sets. We present algebraic structures closely related to partial orders, namely lattices, and Boolean algebras. We study basic issues about lattices, such as their dual roles as special partially ordered sets and algebraic structures, the theory of complete lattices and Galois connections, and their applications to the study of association rules. Special attention is paid to Boolean algebras which are of increasing interest for data mining because they allow the discovery of minimal sets of features necessary for explaining observations and the discovery of hidden patterns.

An introduction to topology and measure theory, which is essential for the study of various concepts of dimension and the recent preoccupations of data mining researchers with the applications of fractal theory to data mining, is also a component of this part.

A variety of applications in data mining are discussed, such as the notion of entropy, presented in a new algebraic framework related to partitions rather than random distributions, levelwise algorithms that generalize the Apriori technique, and generalized measures and their use in the study of frequent item sets. This part concludes with a chapter on rough sets.

The third part is focused on metric spaces. Metrics play an important role in clustering, classification, and certain data preprocessing techniques. We study a variety of concepts related to metrics, from dissimilarities to metrics, tree metrics, and ultrametrics. This chapter is followed by an application chapter dedicated to clustering that includes basic types of clustering algorithms, limitations of clustering, and techniques for evaluating cluster quality.

The fourth part focuses on combinatorics, an area of mathematics dedicated to the study of finite collections of objects that satisfy certain criteria. The main topics discussed are the inclusion-exclusion principle, combinatorics of partitions, counting problems related to collections of sets, and the Vapnik-Chervonenkis dimension of collections of sets.

Each chapter ends with suggestions for further reading. The book contains more than 400 exercises; they form an integral part of the material. Some of the exercises are in reality supplemental material. For these, we include solutions. The mathematics required for making the best use of our book is a typical three-semester sequence in calculus.

We would like to thank Catherine Brett and Frank Ganz from Springer-Verlag for their professionalism and helpfulness.

Boston and Villeneuve d'Ascq — *Dan A. Simovici*
January 2008 — *Chabane Djeraba*

# Contents

**Preface** ..... v

**Part I Set Theory**

**1 Sets, Relations, and Functions** ..... 3
1.1 Introduction ..... 3
1.2 Sets and Collections ..... 3
1.3 Relations and Functions ..... 9
1.3.1 Cartesian Products of Sets ..... 9
1.3.2 Relations ..... 10
1.3.3 Functions ..... 15
1.3.4 Finite and Infinite Sets ..... 22
1.3.5 Generalized Set Products and Sequences ..... 24
1.3.6 Equivalence Relations ..... 30
1.3.7 Partitions and Covers ..... 32
1.4 The Axiom of Choice ..... 34
1.5 Countable Sets ..... 35
1.6 Elementary Combinatorics ..... 38
1.7 Multisets ..... 44
1.8 Relational Databases ..... 46
Exercises and Supplements ..... 49
Bibliographical Comments ..... 55

**2 Algebras** ..... 57
2.1 Introduction ..... 57
2.2 Operations and Algebras ..... 57
2.3 Morphisms, Congruences, and Subalgebras ..... 61
2.4 Linear Spaces ..... 64
2.5 Matrices ..... 68

Exercises and Supplements . . . . . 74
Bibliographical Comments . . . . . 77

**3 Graphs and Hypergraphs** . . . . . 79
3.1 Introduction . . . . . 79
3.2 Basic Notions of Graph Theory . . . . . 79
3.2.1 Degrees of Vertices . . . . . 80
3.2.2 Graph Representations . . . . . 84
3.2.3 Paths . . . . . 85
3.2.4 Directed Graphs . . . . . 86
3.3 Trees . . . . . 92
3.4 Flows in Digraphs . . . . . 111
3.5 Hypergraphs . . . . . 118
Exercises and Supplements . . . . . 121
Bibliographical Comments . . . . . 124

**Part II Partial Orders**

**4 Partially Ordered Sets** . . . . . 129
4.1 Introduction . . . . . 129
4.2 Partial Orders . . . . . 129
4.3 Special Elements of Partially Ordered Sets . . . . . 133
4.4 The Poset of Real Numbers . . . . . 137
4.5 Closure and Interior Systems . . . . . 139
4.6 The Poset of Partitions of a Set . . . . . 144
4.7 Chains and Antichains . . . . . 148
4.8 Poset Product . . . . . 155
4.9 Functions and Posets . . . . . 158
4.10 Posets and the Axiom of Choice . . . . . 160
4.11 Locally Finite Posets and Möbius Functions . . . . . 162
Exercises and Supplements . . . . . 168
Bibliographical Comments . . . . . 172

**5 Lattices and Boolean Algebras** . . . . . 173
5.1 Introduction . . . . . 173
5.2 Lattices as Partially Ordered Sets and Algebras . . . . . 173
5.3 Special Classes of Lattices . . . . . 180
5.4 Complete Lattices . . . . . 188
5.5 Boolean Algebras and Boolean Functions . . . . . 192
5.6 Logical Data Analysis . . . . . 211
Exercises and Supplements . . . . . 219
Bibliographical Comments . . . . . 224

**6 Topologies and Measures** . . . . . 225
6.1 Introduction . . . . . 225
6.2 Topologies . . . . . 225
6.3 Closure and Interior Operators in Topological Spaces . . . . . 226
6.4 Bases . . . . . 235
6.5 Compactness . . . . . 239
6.6 Continuous Functions . . . . . 241
6.7 Connected Topological Spaces . . . . . 244
6.8 Separation Hierarchy of Topological Spaces . . . . . 247
6.9 Products of Topological Spaces . . . . . 249
6.10 Fields of Sets . . . . . 251
6.11 Measures . . . . . 256
Exercises and Supplements . . . . . 265
Bibliographical Comments . . . . . 272

**7 Frequent Item Sets and Association Rules** . . . . . 273
7.1 Introduction . . . . . 273
7.2 Frequent Item Sets . . . . . 273
7.3 Borders of Collections of Sets . . . . . 279
7.4 Association Rules . . . . . 281
7.5 Levelwise Algorithms and Posets . . . . . 283
7.6 Lattices and Frequent Item Sets . . . . . 288
Exercises and Supplements . . . . . 290
Bibliographical Comments . . . . . 292

**8 Applications to Databases and Data Mining** . . . . . 295
8.1 Introduction . . . . . 295
8.2 Tables and Indiscernibility Relations . . . . . 295
8.3 Partitions and Functional Dependencies . . . . . 298
8.4 Partition Entropy . . . . . 305
8.5 Generalized Measures and Data Mining . . . . . 321
8.6 Differential Constraints . . . . . 325
Exercises and Supplements . . . . . 330
Bibliographical Comments . . . . . 332

**9 Rough Sets** . . . . . 333
9.1 Introduction . . . . . 333
9.2 Approximation Spaces . . . . . 333
9.3 Decision Systems and Decision Trees . . . . . 337
9.4 Closure Operators and Rough Sets . . . . . 345
Exercises and Supplements . . . . . 347
Bibliographical Comments . . . . . 348

**Part III Metric Spaces**

**10 Dissimilarities, Metrics, and Ultrametrics** . . . . . 351
10.1 Introduction . . . . . 351
10.2 Classes of Dissimilarities . . . . . 351
10.3 Tree Metrics . . . . . 357
10.4 Ultrametric Spaces . . . . . 366
10.5 Metrics on $\mathbb{R}^n$ . . . . . 377
10.6 Metrics on Collections of Sets . . . . . 388
10.7 Metrics on Partitions . . . . . 394
10.8 Metrics on Sequences . . . . . 398
10.9 Searches in Metric Spaces . . . . . 402
Exercises and Supplements . . . . . 411
Bibliographical Comments . . . . . 421

**11 Topologies and Measures on Metric Spaces** . . . . . 423
11.1 Introduction . . . . . 423
11.2 Metric Space Topologies . . . . . 423
11.3 Continuous Functions in Metric Spaces . . . . . 426
11.4 Separation Properties of Metric Spaces . . . . . 427
11.5 Sequences in Metric Spaces . . . . . 435
11.6 Completeness of Metric Spaces . . . . . 439
11.7 Contractions and Fixed Points . . . . . 445
11.8 Measures in Metric Spaces . . . . . 449
11.9 Embeddings of Metric Spaces . . . . . 452
Exercises and Supplements . . . . . 454
Bibliographical Comments . . . . . 458

**12 Dimensions of Metric Spaces** . . . . . 459
12.1 Introduction . . . . . 459
12.2 The Dimensionality Curse . . . . . 459
12.3 Inductive Dimensions of Topological Metric Spaces . . . . . 462
12.4 The Covering Dimension . . . . . 472
12.5 The Cantor Set . . . . . 475
12.6 The Box-Counting Dimension . . . . . 479
12.7 The Hausdorff-Besicovitch Dimension . . . . . 482
12.8 Similarity Dimension . . . . . 486
Exercises and Supplements . . . . . 490
Bibliographical Comments . . . . . 493

**13 Clustering** . . . . . 495
13.1 Introduction . . . . . 495
13.2 Hierarchical Clustering . . . . . 496
13.2.1 Matrix-Based Hierarchical Clustering . . . . . 498
13.2.2 Graph-based Hierarchical Clustering . . . . . 506
13.3 The $k$-Means Algorithm . . . . . 512
13.4 The PAM Algorithm . . . . . 514
13.5 Limitations of Clustering . . . . . 516
13.6 Clustering Quality . . . . . 520
13.6.1 Object Silhouettes . . . . . 520
13.6.2 Supervised Evaluation . . . . . 521
Exercises and Supplements . . . . . 523
Bibliographical Comments . . . . . 525

**Part IV Combinatorics**

**14 Combinatorics** . . . . . 529
14.1 Introduction . . . . . 529
14.2 The Inclusion-Exclusion Principle . . . . . 529
14.3 Ramsey's Theorem . . . . . 533
14.4 Combinatorics of Partitions . . . . . 536
14.5 Combinatorics of Collections of Sets . . . . . 539
Exercises and Supplements . . . . . 544
Bibliographical Comments . . . . . 549

**15 The Vapnik-Chervonenkis Dimension** . . . . . 551
15.1 Introduction . . . . . 551
15.2 The Vapnik-Chervonenkis Dimension . . . . . 551
15.3 Perceptrons . . . . . 563
Exercises and Supplements . . . . . 565
Bibliographical Comments . . . . . 567

**Part V Appendices**

**A Asymptotics** . . . . . 571

**B Convex Sets and Functions** . . . . . 573

**C Useful Integrals and Formulas** . . . . . 583
C.1 Euler's Integrals . . . . . 583
C.2 Wallis's Formula . . . . . 587
C.3 Stirling's Formula . . . . . 588
C.4 The Volume of an $n$-Dimensional Sphere . . . . . 590

**D A Characterization of a Function** . . . . . 593

**References** . . . . . 597

**Topic Index** . . . . . 605

# Part I

# Set Theory

# 1

# Sets, Relations, and Functions

## 1.1 Introduction

In this chapter, dedicated to set-theoretical bases of data mining, we assume that the reader is familiar with the notion of a set, membership of an element in a set, and elementary set theory. After a brief review of set-theoretical operations we discuss collections of sets, ordered pairs, and set products.

The Axiom of Choice, a basic principle used in many branches of mathematics, is discussed in Section 1.4. This subject is approached again in the context of partially ordered sets in Chapter 4. Countable and uncountable sets are presented in Section 1.5. An introductory section on elementary combinatorics is expanded in Chapter 14. Finally, we introduce the basics of the relational database model.

## 1.2 Sets and Collections

If $x$ is a member of a set $S$, this is denoted, as usual, by $x \in S$. To denote that $x$ is not a member of the set $S$, we write $x \notin S$.

Throughout this book, we use standardized notations for certain important sets of numbers:

| | |
|---|---|
| $\mathbb{C}$ | the set of complex numbers |
| $\mathbb{R}$ | the set of real numbers |
| $\mathbb{R}_{\geq 0}$ | the set of nonnegative real numbers |
| $\mathbb{R}_{>0}$ | the set of positive real numbers |
| $\hat{\mathbb{R}}_{\geq 0}$ | the set $\mathbb{R}_{\geq 0} \cup \{+\infty\}$ |
| $\hat{\mathbb{R}}$ | the set $\mathbb{R} \cup \{-\infty, +\infty\}$ |
| $\mathbb{Q}$ | the set of rational numbers |
| $\mathbb{I}$ | the set of irrational numbers |
| $\mathbb{Z}$ | the set of integers |
| $\mathbb{N}$ | the set of natural numbers |
| $\mathbb{N}_1$ | the set of positive natural numbers |

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_1, © Springer-Verlag London Limited 2008

The usual order of real numbers is extended to the set $\hat{\mathbb{R}}$ by $-\infty < x < +\infty$ for every $x \in \mathbb{R}$. In addition, we assume that

$$x + \infty = \infty + x = +\infty,$$
$$x - \infty = -\infty + x = -\infty,$$

for every $x \in \mathbb{R}$. Also,

$$x \cdot \infty = \infty \cdot x = \begin{cases} +\infty & \text{if } x > 0 \\ -\infty & \text{if } x < 0, \end{cases}$$

and

$$x \cdot (-\infty) = (-\infty) \cdot x = \begin{cases} -\infty & \text{if } x > 0 \\ \infty & \text{if } x < 0. \end{cases}$$

Note that the product of 0 with either $+\infty$ or $-\infty$ is not defined. Division is extended by $x/+\infty = x/-\infty = 0$ for every $x \in \mathbb{R}$.

If $S$ is a finite set, we denote by $|S|$ the number of elements of $S$.

Sets may contain other sets as elements. For example, the set

$$\mathcal{C} = \{\emptyset, \{0\}, \{0,1\}, \{0,2\}, \{1,2,3\}\}$$

contains the empty set $\emptyset$ and $\{0\}, \{0,1\}$,$\{0,2\}$,$\{1,2,3\}$ as its elements. We refer to such sets as *collections of sets* or simply *collections*. In general, we use calligraphic letters $\mathcal{C}, \mathcal{D}, \ldots$ to denote collections of sets.

If $\mathcal{C}$ and $\mathcal{D}$ are two collections, we say that $\mathcal{C}$ is *included* in $\mathcal{D}$, or that $\mathcal{C}$ is a *subcollection* of $\mathcal{D}$, if every member of $\mathcal{C}$ is a member of $\mathcal{D}$. This is denoted by $\mathcal{C} \subseteq \mathcal{D}$.

Two collections $\mathcal{C}$ and $\mathcal{D}$ are equal if we have both $\mathcal{C} \subseteq \mathcal{D}$ and $\mathcal{D} \subseteq \mathcal{C}$. This is denoted by $\mathcal{C} = \mathcal{D}$.

**Definition 1.1.** *Let $\mathcal{C}$ be a collection of sets. The* union of $\mathcal{C}$*, denoted by $\bigcup \mathcal{C}$, is the set defined by*

$$\bigcup \mathcal{C} = \{x \mid x \in S \text{ for some } S \in \mathcal{C}\}.$$

*If $\mathcal{C}$ is a* nonempty *collection, its* intersection *is the set $\bigcap \mathcal{C}$ given by*

$$\bigcap \mathcal{C} = \{x \mid x \in S \text{ for every } S \in \mathcal{C}\}.$$

If $\mathcal{C} = \{S, T\}$, we have $x \in \bigcup \mathcal{C}$ if and only if $x \in S$ or $x \in T$ and $x \in \bigcup \mathcal{C}$ if and only if $x \in S$ and $y \in T$. The union and the intersection of this two-set collection are denoted by $S \cup T$ and $S \cap T$ and are referred to as the union and the intersection of $S$ and $T$, respectively.

We give, without proof, several properties of union and intersection of sets:

1. $S \cup (T \cup U) = (S \cup T) \cup U$ (*associativity of union*),
2. $S \cup T = T \cup S$ (*commutativity of union*),
3. $S \cup S = S$ (*idempotency of union*),
4. $S \cup \emptyset = S$,
5. $S \cap (T \cap U) = (S \cap T) \cap U$ (*associativity of intersection*),
6. $S \cap T = T \cap S$ (*commutativity of intersection*),
7. $S \cap S = S$ (*idempotency of intersection*),
8. $S \cap \emptyset = \emptyset$,

for all sets $S, T, U$.

The associativity of union and intersection allows us to denote unambiguously the union of three sets $S, T, U$ by $S \cup T \cup U$ and the intersection of three sets $S, T, U$ by $S \cup T \cup U$.

**Definition 1.2.** *The sets $S$ and $T$ are* disjoint *if $S \cap T = \emptyset$.*

*A collection of sets $\mathcal{C}$ is said to be a* collection of pairwise disjoint sets *if for every $S$ and $T$ in $\mathcal{C}$, if $S \neq T$, $S$ and $T$ are disjoint.*

**Definition 1.3.** *Let $S$ and $T$ be two sets. The* difference *of $S$ and $T$ is the set $S - T$ defined by*

$$S - T = \{x \in S \mid x \notin T\}.$$

When the set $S$ is understood from the context, we write $\overline{T}$ for $S - T$, and we refer to the set $\overline{T}$ as the *complement* of $T$ with respect to $S$ or simply the *complement of $T$*.

The relationship between set difference and set union and intersection is given in the following theorem.

**Theorem 1.4.** *For every set $S$ and nonempty collection $\mathcal{C}$ of sets, we have*

$$S - \bigcup \mathcal{C} = \bigcap \{S - C \mid C \in \mathcal{C}\},$$
$$S - \bigcap \mathcal{C} = \bigcup \{S - C \mid C \in \mathcal{C}\}.$$

*Proof.* We leave the proof of these equalities to the reader. □

**Corollary 1.5.** *For any sets $S, T, U$, we have*

$$S - (T \cup U) = (S - T) \cap (S - U),$$
$$S - (T \cap U) = (S - T) \cup (S - U).$$

*Proof.* The corollary follows immediately from Theorem 1.4 by choosing $\mathcal{C} = \{T, U\}$. □

With the notation previously introduced for the complement of a set, the equalities of Corollary 1.5 become

$$\overline{T \cup U} = \overline{T} \cap \overline{U},$$
$$\overline{T \cap U} = \overline{T} \cup \overline{U}.$$

The link between union and intersection is given by the distributivity properties contained in the following theorem.

**Theorem 1.6.** *For any collection of sets* $\mathcal{C}$ *and set* $T$*, we have*

$$\left(\bigcup \mathcal{C}\right) \cap T = \bigcup \{C \cap T \mid C \in \mathcal{C}\}.$$

*If* $\mathcal{C}$ *is nonempty, we also have*

$$\left(\bigcap \mathcal{C}\right) \cup T = \bigcap \{C \cup T \mid C \in \mathcal{C}\}.$$

*Proof.* We shall prove only the first equality; the proof of the second one is left as an exercise for the reader.

Let $x \in (\bigcup \mathcal{C}) \cap T$. This means that $x \in \bigcup \mathcal{C}$ and $x \in T$. There is a set $C \in \mathcal{C}$ such that $x \in C$; hence, $x \in C \cap T$, which implies $x \in \bigcup \{C \cap T \mid C \in \mathcal{C}\}$.

Conversely, if $x \in \bigcup \{C \cap T \mid C \in \mathcal{C}\}$, there exists a member $C \cap T$ of this collection such that $x \in C \cap T$, so $x \in C$ and $x \in T$. It follows that $x \in \bigcup \mathcal{C}$, and this, in turn, gives $x \in (\bigcup \mathcal{C}) \cap T$. □

**Corollary 1.7.** *For any sets* $T$*,* $U$*,* $V$*, we have*

$$(U \cup V) \cap T = (U \cap T) \cup (V \cap T),$$
$$(U \cap V) \cup T = (U \cup T) \cap (V \cup T).$$

*Proof.* The corollary follows immediately by choosing $\mathcal{C} = \{U, V\}$ in Theorem 1.6. □

Note that if $\mathcal{C}$ and $\mathcal{D}$ are two collections such that $\mathcal{C} \subseteq \mathcal{D}$, then

$$\bigcup \mathcal{C} \subseteq \bigcup \mathcal{D}$$

and

$$\bigcap \mathcal{D} \subseteq \bigcap \mathcal{C}.$$

We initially excluded the empty collection from the definition of the intersection of a collection. However, within the framework of collections of subsets of a given set $S$, we will extend the previous definition by taking $\bigcap \emptyset = S$ for the empty collection of subsets of $S$. This is consistent with the fact that $\emptyset \subseteq \mathcal{C}$ implies $\bigcap \mathcal{C} \subseteq S$.

The *symmetric difference* of sets denoted by $\oplus$ is defined by

$$U \oplus V = (U - V) \cup (V - U)$$

for all sets $U, V$.

**Theorem 1.8.** *For all sets $U, V, T$, we have*
(i) $U \oplus U = \emptyset$;
(ii) $U \oplus V = V \oplus T$;
(iii) $(U \oplus V) \oplus T = U \oplus (V \oplus T)$.

*Proof.* The first two parts of the theorem are direct applications of the definition of $\oplus$. We leave to the reader the proof of the third part (the associativity of $\oplus$).

The next theorem allows us to introduce a type of set collection of fundamental importance.

**Theorem 1.9.** *Let $\{\{x, y\}, \{x\}\}$ and $\{\{u, v\}, \{u\}\}$ be two collections such that $\{\{x, y\}, \{x\}\} = \{\{u, v\}, \{u\}\}$. Then, we have $x = u$ and $y = v$.*

*Proof.* Suppose that $\{\{x, y\}, \{x\}\} = \{\{u, v\}, \{u\}\}$.

If $x = y$, the collection $\{\{x, y\}, \{x\}\}$ consists of a single set, $\{x\}$, so the collection $\{\{u, v\}, \{u\}\}$ will also consist of a single set. This means that $\{u, v\} = \{u\}$, which implies $u = v$. Therefore, $x = u$, which gives the desired conclusion because we also have $y = v$.

If $x \neq y$, then neither $(x, y)$ nor $(u, v)$ are singletons. However, they both contain exactly one singleton, namely $\{x\}$ and $\{u\}$, respectively, so $x = u$. They also contain the equal sets $\{x, y\}$ and $\{u, v\}$, which must be equal. Since $v \in \{x, y\}$ and $v \neq u = x$, we conclude that $v = y$. □

**Definition 1.10.** *An* ordered pair *is a collection of sets $\{\{x, y\}, \{x\}\}$.*

Theorem 1.9 implies that for an ordered pair $\{\{x, y\}, \{x\}\}$, $x$ and $y$ are uniquely determined. This justifies the following definition.

**Definition 1.11.** *Let $\{\{x, y\}, \{x\}\}$ be an ordered pair. Then $x$ is* the first component of *$p$ and $y$ is* the second component of *$p$.*

From now on, an ordered pair $\{\{x, y\}, \{x\}\}$ will be denoted by $(x, y)$. If both $x, y \in S$, we refer to $(x, y)$ as an *ordered pair on the set $S$.*

**Definition 1.12.** *Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of sets such that $\bigcup \mathcal{C} = \bigcup \mathcal{D}$. $\mathcal{D}$ is a* refinement *of $\mathcal{C}$ if, for every $D \in \mathcal{D}$, there exists $C \in \mathcal{C}$ such that $D \subseteq C$.*
*This is denoted by $\mathcal{C} \sqsubseteq \mathcal{D}$.*

*Example 1.13.* Consider the collection $\mathcal{C} = \{(a, \infty) \mid a \in \mathbb{R}\}$ and $\mathcal{D} = \{(a, b) \mid a, b \in \mathbb{R}, a < b\}$. It is clear that $\bigcup \mathcal{C} = \bigcup \mathcal{D} = \mathbb{R}$.

Since we have $(a, b) \subseteq (a, \infty)$ for every $a, b \in \mathbb{R}$ such that $a < b$, it follows that $\mathcal{D}$ is a refinement of $\mathcal{C}$.

**Definition 1.14.** *A collection of sets $\mathcal{C}$ is* hereditary *if $U \in \mathcal{C}$ and $W \subseteq U$ implies $W \in \mathcal{C}$.*

*Example 1.15.* Let $S$ be a set. The collection of subsets of $S$, denoted by $\mathcal{P}(S)$, is a hereditary collection of sets since a subset of a subset $T$ of $S$ is itself a subset of $S$.

The set of subsets of $S$ that contain $k$ elements is denoted by $\mathcal{P}_k(S)$. Clearly, for every set $S$, we have $\mathcal{P}_0(S) = \{\emptyset\}$ because there is only one subset of $S$ that contains 0 elements, namely the empty set. The set of all finite subsets of a set $S$ is denoted by $\mathcal{P}_{fin}(S)$. It is clear that $\mathcal{P}_{fin}(S) = \bigcup k \in \mathbb{N}\mathcal{P}_k(S)$.

*Example 1.16.* If $S = \{a, b, c\}$, then $\mathcal{P}(S)$ consists of the following eight sets:

$$\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}.$$

For the empty set, we have $\mathcal{P}(\emptyset) = \{\emptyset\}$.

**Definition 1.17.** *A collection $\mathcal{C}$ has* finite character *if $C \in \mathcal{C}$ if and only if every finite subset of $C$ belongs to $\mathcal{C}$.*

It is clear that, for a collection $\mathcal{C}$ of finite character, if $C \in \mathcal{C}$ and $D \subseteq C$, then we also have $D \in \mathcal{C}$. In other words, every collection of finite character is hereditary.

**Theorem 1.18.** *Let $\mathcal{C}$ be a collection of finite character that consists of subsets of a set $S$. If $U_0, \ldots, U_n, \ldots$ are members of $\mathcal{C}$ such that $U_0 \subseteq \cdots \subseteq U_n \subseteq \cdots$, then $U = \bigcup\{U_i \mid i \geq 0\} \in \mathcal{C}$.*

*Proof.* Let $W = \{w_i \mid 0 \leq i \leq n-1\}$ be a finite subset of $U$. For every $w_\ell \in W$, let $w_\ell$ be the least integer such that $w_\ell \in U_{q_\ell}$ for $0 \leq \ell \leq n-1$. If $q = \max\{q_0, \ldots, q_{k-1}\}$, then $W \subseteq U_q$, so $W \in \mathcal{C}$. Since every finite subset of $U$ belongs to $\mathcal{C}$, we obtain $U \in \mathcal{C}$. □

**Definition 1.19.** *Let $\mathcal{C}$ be a collection of sets and let $K$ be a set. The* trace of the collection $\mathcal{C}$ on the set $K$ *is the collection $\{C \cap K \mid C \in \mathcal{C}\}$.*

An alternative notation for $\mathcal{C}_K$ is $\mathcal{C} \restriction_K$, a notation that we shall use when the collection $\mathcal{C}$ is adorned by other subscripts.

We conclude this presentation of collections of sets with two more operations on collections of sets.

**Definition 1.20.** *Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of sets. The collections $\mathcal{C} \vee \mathcal{D}$, $\mathcal{C} \wedge \mathcal{D}$, and $\mathcal{C} - \mathcal{D}$ are given by*

$$\begin{aligned}
\mathcal{C} \vee \mathcal{D} &= \{C \cup D \mid C \in \mathcal{C} \text{ and } D \in \mathcal{D}\},\\
\mathcal{C} \wedge \mathcal{D} &= \{C \cap D \mid C \in \mathcal{C} \text{ and } D \in \mathcal{D}\},\\
\mathcal{C} - \mathcal{D} &= \{C - D \mid C \in \mathcal{C} \text{ and } D \in \mathcal{D}\}.
\end{aligned}$$

*Example 1.21.* Let $\mathcal{C}$ and $\mathcal{D}$ be the collections of sets defined by

$$\mathcal{C} = \{\{x\}, \{y, z\}, \{x, y\}, \{x, y, z\}\},$$
$$\mathcal{D} = \{\{y\}, \{x, y\}, \{u, y, z\}\}.$$

We have

$$\mathcal{C} \vee \mathcal{D} = \{\{x, y\}, \{y, z\}, \{x, y, z\}, \{u, y, z\}, \{u, x, y, z\}\},$$
$$\mathcal{C} \wedge \mathcal{D} = \{\emptyset, \{x\}, \{y\}, \{x, y\}, \{y, z\}\},$$
$$\mathcal{C} - \mathcal{D} = \{\emptyset, \{x\}, \{z\}, \{x, z\}\},$$
$$\mathcal{D} - \mathcal{C} = \{\emptyset, \{u\}, \{x\}, \{y\}, \{u, z\}, \{u, y, z\}\}.$$

Unlike "$\cup$" and "$\cap$", the operations "$\vee$" and "$\wedge$" between collections of sets are not idempotent. Indeed, we have, for example,

$$\mathcal{D} \vee \mathcal{D} = \{\{y\}, \{x, y\}, \{u, y, z\}, \{u, x, y, z\}\} \neq \mathcal{D}.$$

The trace $\mathcal{C}_K$ of a collection $\mathcal{C}$ on $K$ can be written as $\mathcal{C}_K = \mathcal{C} \wedge \{K\}$.

## 1.3 Relations and Functions

This section covers a number of topics that are derived from the notion of relation.

### 1.3.1 Cartesian Products of Sets

**Definition 1.22.** *Let $X$ and $Y$ be two sets. The* Cartesian product *of $X$ and $Y$ is the set $X \times Y$, which consists of all pairs $(x, y)$ such that $x \in X$ and $y \in Y$.*

If either $X = \emptyset$ or $Y = \emptyset$, then $X \times Y = \emptyset$.

*Example 1.23.* Consider the sets $X = \{a, b, c\}$ and $Y = \{0, 1\}$. Their Cartesian product is the set:

$$X \times Y = \{(x, 0), (y, 0), (z, 0), (x, 1), (y, 1), (z, 1)\}.$$

*Example 1.24.* The Cartesian product $\mathbb{R} \times \mathbb{R}$ consists of all ordered pairs of real numbers $(x, y)$. Geometrically, each such ordered pair corresponds to a point in a plane equipped with a system of coordinates. Namely, the pair $(u, v) \in \mathbb{R} \times \mathbb{R}$ is represented by the point $P$ whose $x$-coordinate is $u$ and $y$-coordinate is $v$ (see Figure 1.1)

The Cartesian product is distributive over union, intersection, and difference of sets.

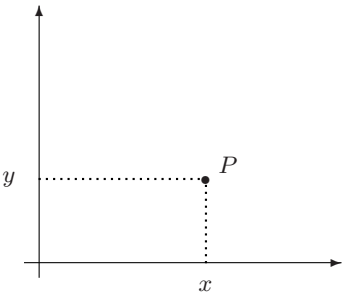



**Fig. 1.1.** Cartesian representation of the pair $(x, y)$.

**Theorem 1.25.** *If $\star$ is one of $\cup, \cap$, or $-$, then for any sets $R$, $S$, and $T$, we have*

$$(R \star S) \times T = (R \times T) \star (S \times T),$$
$$T \times (R \star S) = (T \times R) \star (T \times S).$$

*Proof.* We prove only that $(R - S) \times T = (R \times T) - (S \times T)$. Let $(x, y) \in (R - S) \times T$. We have $x \in R - S$ and $y \in T$. Therefore, $(x, y) \in R \times T$ and $(x, y) \notin S \times T$, which show that $(x, y) \in (R \times T) - (S \times T)$.

Conversely, $(x, y) \in (R \times T) - (S \times T)$ implies $x \in$ and $y \in T$ and also $(x, y) \notin S \times T$. Thus, we have $x \notin S$, so $(x, y) \in (R - S) \times T$. □

It is not difficult to see that if $R \subseteq R'$ and $S \subseteq S'$, then $R \times S \subseteq R' \times S'$. We refer to this property as the *monotonicity of the Cartesian product with respect to set inclusion.*

### 1.3.2 Relations

**Definition 1.26.** *A* relation *is a set of ordered pairs.*

*If $S$ and $T$ are sets and $\rho$ is a relation such that $\rho \subseteq S \times T$, then we refer to $\rho$ as a* relation from $S$ to $T$.

*A relation from $S$ to $S$ is called a* relation on $S$.

$\mathcal{P}(S \times T)$ is the set of all relations from $S$ to $T$.

Among the relations from $S$ to $T$, we distinguish the *empty relation* $\emptyset$ and the *full relation* $S \times T$.

The *identity relation* of a set $S$ is the relation $\iota_S \subseteq S \times S$ defined by $\iota_S = \{(x, x) \mid x \in S\}$. The *full relation on $S$* is $\theta_S = S \times S$.

If $(x, y) \in \rho$, we sometimes denote this fact by $x \,\rho\, y$, and we write $x \not\rho\, y$ instead of $(x, y) \notin \rho$.

*Example 1.27.* Let $S \subseteq \mathbb{R}$. The relation "less than" on $S$ is given by

$$\{(x, y) \mid x, y \in S \text{ and } y = x + z \text{ for some } z \in \mathbb{R}_{\geq 0}\}.$$

*Example 1.28.* Consider the relation $\nu \subseteq \mathbb{Z} \times \mathbb{Q}$ given by

$$\nu = \{(n, q) \mid n \in \mathbb{Z},\ q \in \mathbb{Q},\ \text{and } n \leq q < n + 1\}.$$

We have $(-3, -2.3) \in \nu$ and $(2, 2.3) \in \nu$. Clearly, $(n, q) \in \nu$ if and only if $n$ is the integral part of the rational number $q$.

*Example 1.29.* The relation $\delta$ is defined by

$$\delta = \{(m, n) \in \mathbb{N} \times \mathbb{N} \mid n = km \text{ for some } k \in \mathbb{N}\}.$$

We have $(m, n) \in \delta$ if $m$ divides $n$ evenly.

Note that if $S \subseteq T$, then $\iota_S \subseteq \iota_T$ and $\theta_S \subseteq \theta_T$.

**Definition 1.30.** *The* domain *of a relation $\rho$ from $S$ to $T$ is the set*

$$\mathrm{Dom}(\rho) = \{x \in S \mid (x, y) \in \rho \text{ for some } y \in T\}.$$

*The* range *of $\rho$ from $S$ to $T$ is the set*

$$\mathrm{Ran}(\rho) = \{y \in T \mid (x, y) \in \rho \text{ for some } x \in S\}.$$

If $\rho$ is a relation and $S$ and $T$ are sets, then $\rho$ *is a relation from $S$ to $T$* if and only if $\mathrm{Dom}(\rho) \subseteq S$ and $\mathrm{Ran}(\rho) \subseteq T$. Clearly, $\rho$ is always a relation from $\mathrm{Dom}(\rho)$ to $\mathrm{Ran}(\rho)$.

If $\rho$ and $\sigma$ are relations and $\rho \subseteq \sigma$, then $\mathrm{Dom}(\rho) \subseteq \mathrm{Dom}(\sigma)$ and $\mathrm{Ran}(\rho) \subseteq \mathrm{Ran}(\sigma)$.

If $\rho$ and $\sigma$ are relations, then so are $\rho \cup \sigma$, $\rho \cap \sigma$, and $\rho - \sigma$, and in fact if $\rho$ and $\sigma$ are both relations from $S$ to $T$, then these relations are also relations from $S$ to $T$.

**Definition 1.31.** *Let $\rho$ be a relation. The* inverse *of $\rho$ is the relation $\rho^{-1}$ given by*

$$\rho^{-1} = \{(y, x) \mid (x, y) \in \rho\}.$$

The proofs of the following simple properties are left to the reader:

(i) $\mathrm{Dom}(\rho^{-1}) = \mathrm{Ran}(\rho)$,
(ii) $\mathrm{Ran}(\rho^{-1}) = \mathrm{Dom}(\rho)$,
(iii) if $\rho$ is a relation from $A$ to $B$, then $\rho^{-1}$ is a relation from $B$ to $A$, and
(iv) $(\rho^{-1})^{-1} = \rho$

for every relation $\rho$. Furthermore, if $\rho$ and $\sigma$ are two relations such that $\rho \subseteq \sigma$, then $\rho^{-1} \subseteq \sigma^{-1}$ (monotonicity of the inverse).

**Definition 1.32.** *Let $\rho$ and $\sigma$ be relations. The* product *of $\rho$ and $\sigma$ is the relation $\rho\sigma$, where*

$$\rho\sigma = \{(x, z) \mid \text{ for some } y,\ (x, y) \in \rho,\ \text{and } (y, z) \in \sigma\}.$$

It is easy to see that $\mathrm{Dom}(\rho\sigma) \subseteq \mathrm{Dom}(\rho)$ and $\mathrm{Ran}(\rho\sigma) \subseteq \mathrm{Ran}(\sigma)$. Further, if $\rho$ is a relation from $A$ to $B$ and $\sigma$ is a relation from $B$ to $C$, then $\rho\sigma$ is a relation from $A$ to $C$.

Several properties of the relation product are given in the following theorem.

**Theorem 1.33.** *Let $\rho_1, \rho_2$, and $\rho_3$ be relations. We have*

(i) $\rho_1(\rho_2\rho_3) = (\rho_1\rho_2)\rho_3$ *(associativity of relation product).*

(ii) $\rho_1(\rho_2 \cup \rho_3) = (\rho_1\rho_2) \cup (\rho_1\rho_3)$ *and* $(\rho_1 \cup \rho_2)\rho_3 = (\rho_1\rho_3) \cup (\rho_2\rho_3)$ *(distributivity of relation product over union).*

(iii) $(\rho_1\rho_2)^{-1} = \rho_2^{-1}\rho_1^{-1}$.

(iv) *If $\rho_2 \subseteq \rho_3$, then $\rho_1\rho_2 \subseteq \rho_1\rho_3$ and $\rho_2\rho_1 \subseteq \rho_3\rho_1$ (monotonicity of relation product).*

(v) *If $S$ and $T$ are any sets, then $\iota_S\rho_1 \subseteq \rho_1$ and $\rho_1\iota_T \subseteq \rho_1$. Further, $\iota_S\rho_1 = \rho_1$ if and only if $\mathrm{Dom}(\rho_1) \subseteq S$, and $\rho_1\iota_T = \rho_1$ if and only if $\mathrm{Ran}(\rho_1) \subseteq T$. (Thus, $\rho_1$ is a relation from $S$ to $T$ if and only if $\iota_S\rho_1 = \rho_1 = \rho_1\iota_T$.)*

*Proof.* We prove (i), (ii), and (iv) and leave the other parts as exercises.

To prove Part (i), let $(a, d) \in \rho_1(\rho_2\rho_3)$. There is a $b$ such that $(a, b) \in \rho_1$ and $(b, d) \in \rho_2\rho_3$. This means that there exists $c$ such that $(b, c) \in \rho_2$ and $(c, d) \in \rho_3$. Therefore, we have $(a, c) \in \rho_1\rho_2$, which implies $(a, d) \in (\rho_1\rho_2)\rho_3$. This shows that $\rho_1(\rho_2\rho_3) \subseteq (\rho_1\rho_2)\rho_3$.

Conversely, let $(a, d) \in (\rho_1\rho_2)\rho_3$. There is a $c$ such that $(a, c) \in \rho_1\rho_2$ and $(c, d) \in \rho_3$. This implies the existence of a $b$ for which $(a, b) \in \rho_1$ and $(b, c) \in \rho_3$. For this $b$, we have $(b, d) \in \rho_2\rho_3$, which gives $(a, d) \in \rho_1(\rho_2\rho_3)$. We have proven the reverse inclusion, $(\rho_1\rho_2)\rho_3 \subseteq \rho_1(\rho_2\rho_3)$, which gives the associativity of relation product.

For Part (ii), let $(a, c) \in \rho_1(\rho_2 \cup \rho_3)$. Then, there is a $b$ such that $(a, b) \in \rho_1$ and $(b, c) \in \rho_2$ or $(b, c) \in \rho_3$. In the first case, we have $(a, c) \in \rho_1\rho_2$; in the second, $(a, c) \in \rho_1\rho_3$. Therefore, we have $(a, c) \in (\rho_1\rho_2) \cup (\rho_1\rho_3)$ in either case, so $\rho_1(\rho_2 \cup \rho_3) \subseteq (\rho_1\rho_2) \cup (\rho_1\rho_3)$.

Let $(a, c) \in (\rho_1\rho_2) \cup (\rho_1\rho_3)$. We have either $(a, c) \in \rho_1\rho_2$ or $(a, c) \in \rho_1\rho_3$. In the first case, there is a $b$ such that $(a, b) \in \rho_1$ and $(b, c) \in \rho_2 \subseteq \rho_2 \cup \rho_3$. Therefore, $(a, c) \in \rho_1(\rho_2 \cup \rho_3)$. The second case is handled similarly. This establishes

$$(\rho_1\rho_2) \cup (\rho_1\rho_3) \subseteq \rho_1(\rho_2 \cup \rho_3).$$

The other distributivity property has a similar argument.

Finally, for Part (iv), let $\rho_2$ and $\rho_3$ be such that $\rho_2 \subseteq \rho_3$. Since $\rho_2 \cup \rho_3 = \rho_3$, we obtain from (ii) that

$$\rho_1\rho_3 = (\rho_1\rho_2) \cup (\rho_1\rho_3),$$

which shows that $\rho_1\rho_2 \subseteq \rho_1\rho_3$. The second inclusion is proven similarly. □

**Definition 1.34.** *The $n$-power of a relation $\rho \subseteq S \times S$ is defined inductively by $\rho^0 = \iota_S$ and $\rho^{n+1} = \rho^n\rho$ for $n \in \mathbb{N}$.*

Note that $\rho^1 = \rho^0\rho = \iota_S\rho = \rho$ for any relation $\rho$.

*Example 1.35.* Let $\rho \subseteq \mathbb{R} \times \mathbb{R}$ be the relation defined by

$$\rho = \{(x, x+1) \mid x \in \mathbb{R}\}.$$

The zero-th power of $\rho$ is the relation $\iota_{\mathbb{R}}$. The second power of $\rho$ is

$$\rho^2 = \rho \cdot \rho = \{(x, y) \in \mathbb{R} \times \mathbb{R} \mid (x, z) \in \rho \text{ and } (z, y) \in \rho \text{ for some } z \in \mathbb{R}\}.$$

In other words, $\rho^2 = \{(x, x+2) \mid x \in \mathbb{R}\}$. In general, $\rho^n = \{(x, x+n) \mid x \in \mathbb{R}\}$.

**Definition 1.36.** *A relation $\rho$ is a* function *if for all $x, y, z$, $(x, y) \in \rho$ and $(x, z) \in \rho$ imply $y = z$; $\rho$ is a* one-to-one relation *if, for all $x, x'$, and $y$, $(x, y) \in \rho$ and $(x', y) \in \rho$ imply $x = x'$.*

Observe that $\emptyset$ is a function (referred to in this context as the *empty function*) because $\emptyset$ satisfies vacuously the defining condition for being a function.

*Example 1.37.* Let $S$ be a set. The relation $\rho$ on $S \times \mathcal{P}(S)$ given by

$$\rho = \{(x, \{x\}) \mid x \in S\}$$

is a function.

*Example 1.38.* For every set $S$, the relation $\iota_S$ is both a function and a one-to-one relation. The relation $\nu$ from Example 1.28 is a one-to-one relation, but it is not a function.

**Theorem 1.39.** *For any relation $\rho$, $\rho$ is a function if and only if $\rho^{-1}$ is a one-to-one relation.*

*Proof.* Suppose that $\rho$ is a function, and let $(y_1, x), (y_2, x) \in \rho^{-1}$. Definition 1.31 implies that $(x, y_1), (x, y_2) \in \rho$; hence, $y_1 = y_2$ because $\rho$ is a function. This proves that $\rho^{-1}$ is one-to-one.

Conversely, assume that $\rho^{-1}$ is one-to-one and let $(x, y_1), (x, y_2) \in \rho$. Applying Definition 1.31, we obtain $(y_1, x), (y_2, x) \in \rho^{-1}$ and, since $\rho^{-1}$ is one-to-one, we have $y_1 = y_2$. This shows that $\rho$ is a function. □

*Example 1.40.* We observed that the relation $\nu$ introduced in Example 1.28 is one-to-one. Therefore, its inverse $\nu^{-1} \subseteq \mathbb{Q} \times \mathbb{Z}$ is a function. In fact, $\nu^{-1}$ associates to each rational number $q$ its integer part $\lfloor q \rfloor$.

**Definition 1.41.** *A relation $\rho$ from $S$ to $T$ is* total *if* $\mathrm{Dom}(\rho) = S$ *and is* onto *if* $\mathrm{Ran}(\rho) = T$.

Any relation $\rho$ is a total and onto relation from $\mathrm{Dom}(\rho)$ to $\mathrm{Ran}(\rho)$. If both $S$ and $T$ are nonempty, then $S \times T$ is a total and onto relation from $S$ to $T$.

It is easy to prove that a relation $\rho$ from $S$ to $T$ is a total relation from $S$ to $T$ if and only if $\rho^{-1}$ is an onto relation from $T$ to $S$.

If $\rho$ is a relation, then one can determine whether or not $\rho$ is a function or is one-to-one just by looking at the ordered pairs of $\rho$. Whether $\rho$ is a total or onto relation from $A$ to $B$ depends on what $A$ and $B$ are.

**Theorem 1.42.** *Let $\rho$ and $\sigma$ be relations.*
(i) *If $\rho$ and $\sigma$ are functions, then $\rho\sigma$ is also a function.*
(ii) *If $\rho$ and $\sigma$ are one-to-one relations, then $\rho\sigma$ is also a one-to-one relation.*
(iii) *If $\rho$ is a total relation from $R$ to $S$ and $\sigma$ is a total relation from $S$ to $T$, then $\rho\sigma$ is a total relation from $R$ to $T$.*
(iv) *If $\rho$ is an onto relation from $R$ to $S$ and $\sigma$ is an onto relation from $S$ to $T$, then $\rho\sigma$ is an onto relation from $R$ to $T$.*

*Proof.* To show Part (i), suppose that $\rho$ and $\sigma$ are both functions and that $(x, z_1)$ and $(x, z_2)$ both belong to $\rho\sigma$. Then, there exists a $y_1$ such that $(x, y_1) \in \rho$ and $(y_1, z_1) \in \sigma$, and there exists a $y_2$ such that $(x, y_2) \in \rho$ and $(y_2, z_2) \in \sigma$. Since $\rho$ is a function, $y_1 = y_2$, and hence, since $\sigma$ is a function, $z_1 = z_2$, as desired.

Part (ii) follows easily from Part (i). Suppose that relations $\rho$ and $\sigma$ are one-to-one (and hence that $\rho^{-1}$ and $\sigma^{-1}$ are both functions). To show that $\rho\sigma$ is one-to-one, it suffices to show that $(\rho\sigma)^{-1} = \sigma^{-1}\rho^{-1}$ is a function. This follows immediately from Part (i).

We leave the proofs for the last two parts of the theorem to the reader. □

The properties of relations defined next allow us to define important classes of relations.

**Definition 1.43.** *Let $S$ be a set and let $\rho \subseteq S \times S$ be a relation. The relation $\rho$ is:*
(i) reflexive *if $(s, s) \in \rho$ for every $s \in S$;*
(ii) irreflexive *if $(s, s) \notin \rho$ for every $s \in S$;*
(iii) symmetric *if $(s, s') \in \rho$ implies $(s', s) \in \rho$ for $s, s' \in S$;*
(iv) antisymmetric *if $(s, s'), (s', s) \in \rho$ implies $s = s'$ for $s, s' \in S$;*
(v) asymmetric *if $(s, s') \in \rho$ implies $(s', s) \notin \rho$; and*
(vi) transitive *if $(s, s'), (s', s'') \in \rho$ implies $(s, s'') \in \rho$.*

*Example 1.44.* The relation $\iota_S$ is reflexive, symmetric, antisymmetric, and transitive for any set $S$.

*Example 1.45.* The relation $\delta$ introduced in Example 1.29 is reflexive since $n \cdot 1 = n$ for any $n \in \mathbb{N}$.

Suppose that $(m, n), (n, m) \in \delta$. There are $p, q \in \mathbb{N}$ such that $mp = n$ and $nq = m$. If $n = 0$, then this also implies $m = 0$; hence, $m = n$. Let us assume

that $n \neq 0$. The previous equalities imply $nqp = n$, and since $n \neq 0$, we have $qp = 1$. In view of the fact that both $p$ and $q$ belong to $\mathbb{N}$, we have $p = q = 1$; hence, $m = n$, which proves the antisymmetry of $\rho$.

Let $(m, n), (n, r) \in \delta$. We can write $n = mp$ and $r = nq$ for some $p, q \in \mathbb{N}$, which gives $r = mpq$. This means that $(m, r) \in \delta$, which shows that $\delta$ is also transitive.

**Definition 1.46.** *Let $S$ and $T$ be two sets and let $\rho \subseteq S \times T$ be a relation.*

*The* image of an element $s \in S$ under the relation $\rho$ *is the set* $\rho(s) = \{t \in T \mid (s, t) \in \rho\}$.

*The* preimage of an element $t \in T$ under $\rho$ *is the set* $\{s \in S \mid (s, t) \in \rho\}$, *which equals* $\rho^{-1}(t)$, *using the previous notation.*

*The* collection of images of $S$ under $\rho$ *is*

$$\mathit{IM}_\rho = \{\rho(s) \mid s \in S\},$$

*while the* collection of preimages of $T$ *is*

$$\mathit{PIM}_\rho = \mathit{IM}_{\rho^{-1}} = \{\rho^{-1}(t) \mid t \in T\}.$$

*If $\mathcal{C}$ and $\mathcal{C}'$ are two collections of subsets of $S$ and $T$, respectively, and $\mathcal{C}' = \mathit{IM}_\rho$ and $\mathcal{C} = \mathit{PIM}_\rho$ for some relation $\rho \subseteq S \times T$, we refer to $\mathcal{C}'$ as the* dual class relative to $\rho$ *of $\mathcal{C}$.*

*Example 1.47.* Any collection $\mathcal{D}$ of subsets of $S$ can be regarded as the collection of images under a suitable relation. Indeed, let $\mathcal{C}$ be such a collection. Define the relation $\rho \subseteq S \times \mathcal{C}$ as $\rho = \{(s, C) \mid s \in S, C \in \mathcal{C} \text{ and } c \in C\}$. Then, $\mathsf{IM}_\rho$ consists of all subsets of $\mathcal{P}(\mathcal{C})$ of the form $\rho(s) = \{C \in \mathcal{C} \mid s \in C\}$ for $s \in S$. It is easy to see that $\mathsf{PIM}_\rho(\mathcal{C}) = \mathcal{C}$.

The collection $\mathsf{IM}_\rho$ defined in this example is referred to as the *bi-dual collection of* $\mathcal{C}$.

### 1.3.3 Functions

We saw that a function is a relation $\rho$ such that, for every $x$ in $\mathrm{Dom}(\rho)$, there is only one $y$ such that $(x, y) \in \rho$. In other words, a function assigns a unique value to each member of its domain.

From now on, we will use the letters $f, g, h$, and $k$ to denote functions, and we will denote the identity relation $\iota_S$, which we have already remarked is a function, by $1_S$.

If $f$ is a function, then, for each $x$ in $\mathrm{Dom}(f)$, we let $f(x)$ denote the unique $y$ with $(x, y) \in f$, and we refer to $f(x)$ as the *image of $x$ under $f$*.

**Definition 1.48.** *Let $S$ and $T$ be sets. A* partial function from $S$ to $T$ *is a relation from $S$ to $T$ that is a function.*

*A* total function from $S$ to $T$ *(also called a* function from $S$ to $T$ *or a* mapping from $S$ to $T$*) is a partial function from $S$ to $T$ that is a total relation from $S$ to $T$.*

The set of all partial functions from $S$ to $T$ is denoted by $S \rightsquigarrow T$ and the set of all total functions from $S$ to $T$ by $S \longrightarrow T$. We have $S \longrightarrow T \subseteq S \rightsquigarrow T$ for all sets $S$ and $T$.

The fact that $f$ is a partial function from $S$ to $T$ is indicated by writing $f : S \rightsquigarrow T$ rather than $f \in S \rightsquigarrow T$. Similarly, instead of writing $f \in S \longrightarrow T$, we use the notation $f : S \longrightarrow T$.

For any sets $S$ and $T$, we have $\emptyset \in S \rightsquigarrow T$. If either $S$ or $T$ is empty, then $\emptyset$ is the only partial function from $S$ to $T$. If $S = \emptyset$, then the empty function is a total function from $S$ to any $T$. Thus, for any sets $S$ and $T$, we have

$$\begin{aligned} S \rightsquigarrow \emptyset &= \{\emptyset\}, \\ \emptyset \rightsquigarrow T &= \{\emptyset\}, \\ \emptyset \longrightarrow T &= \{\emptyset\}. \end{aligned}$$

Furthermore, if $S$ is nonempty, then there can be no (total) function from $S$ to the empty set, so we have

$$S \longrightarrow \emptyset = \emptyset \text{ (if } S \neq \emptyset).$$

**Definition 1.49.** *A one-to-one function is called an* injection.

*A function $f : S \rightsquigarrow T$ is called a* surjection (from $S$ to $T$) *if $f$ is an onto relation from $S$ to $T$, and it is called a* bijection (from $S$ to $T$) *or a* one-to-one correspondence between $S$ and $T$ *if it is total, an injection, and a surjection.*

Using our notation for functions, we can restate the definition of injection as follows: $f$ is an injection if for all $s, s' \in \mathrm{Dom}(f)$, $f(s) = f(s')$ implies $s = s'$. Likewise, $f : S \rightsquigarrow T$ is a surjection if for every $t \in T$ there is an $s \in S$ with $f(s) = t$.

*Example 1.50.* Let $S$ and $T$ be two sets and assume that $S \subseteq T$. The *containment mapping* $c : S \longrightarrow T$ defined by $c(s) = s$ for $s \in S$ is an injection. We denote such a containment by $c : S \hookrightarrow T$.

*Example 1.51.* Let $m \in \mathbb{N}$ be a natural number, $m \geq 2$. Consider the function $r_m : \mathbb{N} \longrightarrow \{0, \ldots, m-1\}$, where $r_m(n)$ is the remainder when $n$ is divided by $m$. Obviously, $r_m$ is well-defined since the remainder $p$ when a natural number is divided by $m$ satisfies $0 \leq p \leq m-1$. The function $r_m$ is onto because of the fact that, for any $p \in \{0, \ldots, m-1\}$, we have $r_m(km + p) = p$ for any $k \in \mathbb{N}$.

For instance, if $m = 4$, we have $r_4(0) = r_4(4) = r_4(8) = \cdots = 0$, $r_4(1) = r_4(5) = r_4(9) = \cdots = 1$, $r_4(2) = r_4(6) = r_4(10) = \cdots = 2$ and $r_4(3) = r_4(7) = r_4(11) = \cdots = 3$.

*Example 1.52.* Let $\mathcal{P}_{fin}(\mathbb{N})$ be the set of finite subsets of $\mathbb{N}$. Define the function $\phi : \mathcal{P}_{fin}(\mathbb{N}) \longrightarrow \mathbb{N}$ as

$$\phi(K) = \begin{cases} 0 & \text{if } K = \emptyset, \\ \sum_{i=1}^{p} 2^{n_i} & \text{if } K = \{n_1, \ldots, n_p\}. \end{cases}$$

It is easy to see that $\phi$ is a bijection.

Since a function is a relation, the ideas introduced in the previous section for relations in general can be equally well applied to functions. In particular, we can consider the inverse of a function and the product of two functions.

If $f$ is a function, then, by Theorem 1.39, $f^{-1}$ is a one-to-one relation; however, $f^{-1}$ is not necessarily a function. In fact, by the same theorem, if $f$ is a function, then $f^{-1}$ is a function if and only if $f$ is an injection.

Suppose now that $f : S \rightsquigarrow T$ is an injection. Then, $f^{-1} : T \rightsquigarrow S$ is also an injection. Further, $f^{-1} : T \rightsquigarrow S$ is total if and only if $f : S \rightsquigarrow T$ is a surjection, and $f^{-1} : T \rightsquigarrow S$ is a surjection if and only if $f : S \rightsquigarrow T$ is total. It follows that $f : S \rightsquigarrow T$ is a bijection if and only if $f^{-1} : T \rightsquigarrow S$ is a bijection.

If $f$ and $g$ are functions, then we will always use the alternative notation $gf$ instead of the notation $fg$ used for the relation product. We will refer to $gf$ as the *composition* of $f$ and $g$ rather than the product.

By Theorem 1.42, the composition of two functions is a function. In fact, it follows from the definition of composition that

$$\mathrm{Dom}(gf) = \{s \in \mathrm{Dom}(f) \mid f(s) \in \mathrm{Dom}(g)\}$$

and, for all $s \in \mathrm{Dom}(gf)$,

$$gf(s) = g(f(s)).$$

This explains why we use $gf$ rather than $fg$. If we used the other notation, the previous equation would become $fg(s) = g(f(s))$, which is rather confusing.

**Definition 1.53.** *Let $f : S \longrightarrow T$. A* left inverse (relative to $S$ and $T$) *for $f$ is a function $g : T \longrightarrow S$ such that $gf = 1_S$. A* right inverse (relative to $S$ and $T$) *for $f$ is a function $g : T \longrightarrow S$ such that $fg = 1_T$.*

**Theorem 1.54.** *Let $f : S \longrightarrow T$.*

(i) *$f$ is a surjection if and only if $f$ has a right inverse (relative to $S$ and $T$).*

(ii) *If $S$ is nonempty, then $f$ is an injection if and only if $f$ has a left inverse (relative to $S$ and $T$).*

*Proof.* To prove the first part, suppose first that $f : S \longrightarrow T$ is a surjection. Define a function $g : T \longrightarrow S$ as follows: For each $y \in T$, let $g(y)$ be some arbitrarily chosen element $x \in S$ such that $f(x) = y$. (Such an $x$ exists because $f$ is surjective.) Then, by definition, $f(g(y)) = y$ for all $y \in T$, so $g$ is a right inverse for $f$. Conversely, suppose that $f$ has a right inverse $g$. Let $y \in T$ and let $x = g(y)$. Then, we have $f(x) = f(g(y)) = 1_T(y) = y$. Thus, $f$ is surjective.

To prove the second part, first suppose that $f : S \longrightarrow T$ is an injection and $S$ is nonempty. Let $x_0$ be some fixed element of $S$. Define a function

$g : T \longrightarrow S$ as follows: If $y \in \text{Ran}(f)$, then, since $f$ is an injection, there is a unique element $x \in S$ such that $f(x) = y$. Define $g(y)$ to be this $x$. If $y \in T - \text{Ran}(f)$, define $g(y) = x_0$. Then, it is immediate from the definition of $g$ that, for all $x \in S$, $g(f(x)) = x$, so $g$ is a left inverse for $f$. Conversely, suppose that $f$ has a left inverse $g$. For all $x_1, x_2 \in S$, if $f(x_1) = f(x_2)$, we have $x_1 = 1_S(x_1) = g(f(x_1)) = g(f(x_2)) = 1_S(x_2) = x_2$. Hence, $f$ is an injection. □

We have used in this proof (without an explicit mention) an axiom of set theory that we discuss in Section 1.4. For a proof that makes explicit use of this axiom, see Supplement 38.

**Theorem 1.55.** *Let $f : S \longrightarrow T$. Then, the following statements are equivalent:*

(i) *$f$ is a bijection.*
(ii) *There is a function $g : T \longrightarrow S$ that is both a left and a right inverse for $f$.*
(iii) *$f$ has both a left inverse and a right inverse.*

*Further, if $f$ is a bijection, then $f^{-1}$ is the only left inverse that $f$ has, and it is the only right inverse that $f$ has.*

*Proof.* (i) implies (ii): If $f : S \longrightarrow B$ is a bijection, then $f^{-1} : T \longrightarrow S$ is both a left and a right inverse for $f$.

(ii) implies (iii): This implication is obvious.

(iii) implies (i): If $f$ has both a left inverse and a right inverse and $S \neq \emptyset$, then it follows immediately from Theorem 1.54 that $f$ is both injective and surjective, so $f$ is a bijection. If $S = \emptyset$, then the existence of a left inverse function from $T$ to $S$ implies that $T$ is also empty; this means that $f$ is the empty function, which is a bijection from the empty set to itself.

Finally, suppose that $f : S \longrightarrow T$ is a bijection and that $g : T \longrightarrow S$ is a left inverse for $f$. Then, we have

$$f^{-1} = 1_S f^{-1} = (gf)f^{-1} = g(ff^{-1}) = g1_T = g.$$

Thus, $f^{-1}$ is the unique left inverse for $f$. A similar proof shows that $f^{-1}$ is the unique right inverse for $f$. □

To prove that $f : S \longrightarrow T$ is a bijection one could prove directly that $f$ is both one-to-one and onto. Theorem 1.55 provides an alternative way. If we can define a function $g : T \longrightarrow S$ and show that $g$ is both a left and a right inverse for $f$, then $f$ is a bijection and $g = f^{-1}$.

The next definition provides another way of viewing a subset of a set $S$.

**Definition 1.56.** *Let $S$ be a set. An* indicator function *over $S$ is a function $I : S \longrightarrow \{0, 1\}$.*

*If $P$ is a subset of $S$, then the* indicator function *of $P$ (as a subset of $S$) is the function $I_P : S \longrightarrow \{0, 1\}$ given by*

$$I_P(x) = \begin{cases} 1 & \text{if } x \in P \\ 0 & \text{otherwise,} \end{cases}$$

*for every* $x \in S$.

It is easy to see that

$$\begin{aligned} I_{P \cap Q}(x) &= I_P(x) \cdot I_Q(x), \\ I_{P \cup Q}(x) &= I_P(x) + I_Q(x) - I_P(x) \cdot I_Q(x), \\ I_{\bar{P}}(x) &= 1 - I_P(x), \end{aligned}$$

for every $P, Q \subseteq S$ and $x \in S$.

The relationship between the subsets of a set and indicator functions defined on that set is discussed next.

**Theorem 1.57.** *There is a bijection* $\Psi : \mathcal{P}(S) \longrightarrow (S \longrightarrow \{0, 1\})$ *between the set of subsets of* $S$ *and the set of indicator functions defined on* $S$.

*Proof.* For $P \in \mathcal{P}(S)$, define $\Psi(P) = I_P$. The mapping $\Psi$ is one-to-one. Indeed, assume that $I_P = I_Q$, where $P, Q \in \mathcal{P}(S)$. We have $x \in P$ if and only if $I_P(x) = 1$, which is equivalent to $I_Q(x) = 1$. This happens if and only if $x \in Q$; hence, $P = Q$ so $\Psi$ is one-to-one.

Let $f : S \longrightarrow \{0, 1\}$ be an arbitrary function. Define the set $T_f = \{x \in S \mid f(x) = 1\}$. It is easy to see that $f$ is the indicator function of the set $T_f$. Hence, $\Psi(T_f) = f$, which shows that the mapping $\Psi$ is also onto and hence it is a bijection. □

**Definition 1.58.** *A* simple function on a set $S$ *is a function* $f : S \longrightarrow \mathbb{R}$ *that has a finite range.*

Simple functions are linear combinations of indicator functions, as we show next.

**Theorem 1.59.** *Let* $f : S \longrightarrow \mathbb{R}$ *be a simple function such that* $\text{Ran}(f) = \{y_1, \ldots, y_n\} \subseteq \mathbb{R}$. *Then,*

$$f = \sum_{i=1}^{n} y_i I_{f^{-1}(y_i)}.$$

*Proof.* Let $x \in \mathbb{R}$. If $f(x) = y_j$, then

$$I_{f^{-1}(y_\ell)}(x) = \begin{cases} 1 & \text{if } \ell = j, \\ 0 & \text{otherwise.} \end{cases}$$

Thus,

$$\left( \sum_{i=1}^{n} y_i I_{f^{-1}(y_i)} \right)(x) = y_j,$$

which shows that $f(x) = \left(\sum_{i=1}^{n} y_i I_{f^{-1}(y_i)}\right)(x)$. □

**Theorem 1.60.** *Let $f_1, \ldots, f_k$ be $k$ simple functions defined on a set $S$. If $g : \mathbb{R}^k \longrightarrow \mathbb{R}$ is an arbitrary function, then $g(f_1, \ldots, f_k)$ is a simple function on $S$ and we have*

$$g(f_1, \ldots, f_k)(x) = \sum_{p_1=1}^{m_1} \cdots \sum_{p_k=1}^{m_k} g(y_{1p_1}, \ldots, y_{kp_k}) I_{f_1^{-1}(y_{1p_1}) \cap \cdots \cap f_k^{-1}(y_{kp_k})}(x)$$

*for every $x \in S$, where $\mathrm{Ran}(f_i) = \{y_{i1}, \ldots, y_{im_i}\}$ for $1 \leq i \leq k$.*

*Proof.* It is clear that the function $g(f_1, \ldots, f_k)$ is a simple function because it has a finite range. Moreover, if $\mathrm{Ran}(f_i) = \{y_{i1}, \ldots, y_{im_i}\}$, then the values of $g(f_1, \ldots, f_k)$ have the form $g(y_{1p_1}, \ldots, y_{kp_k})$, and $g(f_1, \ldots, f_k)$ can be written as

$$\begin{aligned} &g(f_1, \ldots, f_k)(x) \\ &= \sum_{p_1=1}^{m_1} \cdots \sum_{p_k=1}^{m_k} g(y_{1p_1}, \ldots, y_{kp_k}) I_{f_1^{-1}(y_{1p_1})}(x) \cdots I_{f_k^{-1}(y_{kp_k})}(x) \\ &= \sum_{p_1=1}^{m_1} \cdots \sum p_k = 1^{m_k} g(y_{1p_1}, \ldots, y_{kp_k}) I_{f_1^{-1}(y_{1p_1}) \cap \cdots \cap f_k^{-1}(y_{kp_k})}(x) \end{aligned}$$

for $x \in S$. □

Theorem 1.60 justifies the following statement.

**Theorem 1.61.** *If $f_1, \ldots, f_k$ are simple functions on a set $S$, then*

$$\begin{aligned} &\max\{f_1(x), \ldots, f_k(x)\}, \\ &\min\{f_1(x), \ldots, f_k(x)\}, \\ &f_1(x) + \cdots + f_k(x), \\ &f_1(x) \cdot \cdots \cdot f_k(x) \end{aligned}$$

*are simple functions on $S$.*

*Proof.* The statement follows immediately from Theorem 1.60. □

### Functions and Sets

Let $f : S \longrightarrow T$ be a function. If $L$ is a subset of $S$, we define the subset $f(L)$ of $T$ as $f(L) = \{f(s) \mid s \in L\}$. The set $f(L)$ is *the image of $L$ under $f$*.

Also, if $H$ is a subset of $T$, we define the set $f^{-1}(H)$ as the subset of $S$ given by $f^{-1}(H) = \{s \in S \mid f(s) \in H\}$ and we refer to this set as the *inverse image of $H$ under $f$*.

It is easy to verify that $L \subseteq L'$ implies $f(L) \subseteq f(L')$ (*monotonicity of set images*) and $H \subseteq H'$ implies $f^{-1}(H) \subseteq f^{-1}(H')$ for every $L, L' \in \mathcal{P}(S)$ and $H, H' \in \mathcal{P}(T)$ (*monotonicity of set inverse images*).

Next, we discuss the behavior of images and inverse images of sets with respect to union and intersection.

**Theorem 1.62.** *Let $f : S \longrightarrow T$ be a function. If $\mathcal{C}$ is a collection of subsets of $S$, then we have*
(i) $f(\bigcup \mathcal{C}) = \bigcup\{f(L) \mid L \in \mathcal{C}\}$ *and*
(ii) $f(\bigcap \mathcal{C}) \subseteq \bigcap\{f(L) \mid L \in \mathcal{C}\}$.

*Proof.* Note that $L \subseteq \bigcup \mathcal{C}$ for every $L \in \mathcal{C}$. The monotonicity of set images implies $f(L) \subseteq f(\bigcup \mathcal{C})$. Therefore, $\bigcup\{f(L) \mid L \in \mathcal{C}\} \subseteq f(\bigcup \mathcal{C})$.

Conversely, let $t \in f(\bigcup \mathcal{C})$. There is $s \in \bigcup \mathcal{C}$ such that $t = f(s)$. Further, since $s \in \bigcup \mathcal{C}$ we have $s \in L$, for some $L \in \mathcal{C}$, which shows that $f \in f(L) \subseteq \bigcup\{f(L) \mid L \in \mathcal{C}\}$, which implies the reverse inclusion $f(\bigcup \mathcal{C}) \subseteq \bigcup\{f(L) \mid L \in \mathcal{C}\}$.

We leave to the reader the second part of the theorem. □

**Theorem 1.63.** *Let $f : S \longrightarrow T$ and $g : T \longrightarrow U$ be two functions. We have $f^{-1}(g^{-1}(X)) = (gf)^{-1}(X)$ for every subset $X$ of $U$.*

*Proof.* We have $s \in f^{-1}(g^{-1}(X))$ if and only if $f(s) \in g^{-1}(X)$, which is equivalent to $g(f(s)) \in X$, that is, with $s \in (gf)^{-1}(X)$. The equality of the theorem follows immediately. □

**Theorem 1.64.** *If $f : S \longrightarrow T$ is an injective function, then $f(\bigcap \mathcal{C}) = \bigcap\{f(L) \mid L \in \mathcal{C}\}$ for every collection $\mathcal{C}$ of subsets of $S$.*

*Proof.* By Theorem 1.62, it suffices to show that for an injection $f$ we have $\bigcap\{f(L) \mid L \in \mathcal{C}\} \subseteq f(\bigcap \mathcal{C})$.

Let $y \in \bigcap\{f(L) \mid L \in \mathcal{C}\}$. For each set $L \in \mathcal{C}$ there exists $x_L \in L$ such that $f(x_L) = y$. Since $f$ is an injection, it follows that there exists $x \in S$ such that $x_L = x$ for every $L \in \mathcal{C}$. Thus, $x \in \bigcap \mathcal{C}$, which implies that $y = f(x) \in f(\bigcap \mathcal{C})$. This allows us to obtain the desired inclusion. □

**Theorem 1.65.** *Let $f : S \longrightarrow T$ be a function. If $\mathcal{D}$ is a collection of subsets of $T$, then we have*
(i) $f^{-1}(\bigcup \mathcal{D}) = \bigcup\{f^{-1}(H) \mid H \in \mathcal{D}\}$ *and*
(ii) $f^{-1}(\bigcap \mathcal{D}) = \bigcap\{f^{-1}(H) \mid H \in \mathcal{D}\}$.

*Proof.* We prove only the second part of the theorem and leave the first part to the reader.

Since $\bigcap \mathcal{D} \subseteq H$ for every $H \in \mathcal{D}$, we have $f^{-1}(\bigcap \mathcal{D}) \subseteq f^{-1}(H)$ due to the monotonicity of set inverse images. Therefore, $f^{-1}(\bigcap \mathcal{D}) \subseteq \bigcap\{f^{-1}(H) \mid H \in \mathcal{D}\}$.

To prove the reverse inclusion, let $s \in \bigcap\{f^{-1}(H) \mid H \in \mathcal{D}\}$. This means that $s \in f^{-1}(H)$ and therefore $f(s) \in H$ for every $H \in \mathcal{D}$. This implies $f(s) \in \bigcap \mathcal{D}$, so $s \in f^{-1}(\bigcap \mathcal{D})$, which yields the reverse inclusion $\bigcap\{f^{-1}(H) \mid H \in \mathcal{D}\} \subseteq f^{-1}(\bigcap \mathcal{D})$. □

Note that images and inverse images behave differently with respect to intersection. The inclusion contained by the second part of Theorem 1.62 may be strict, as the following example shows.

*Example 1.66.* Let $S = \{s_0, s_1, s_2\}$, $T = \{t_0, t_1\}$, and $f : S \longrightarrow T$ be the function defined by $f(s_0) = f(s_1) = t_0$ and $f(s_2) = t_1$. Consider the collection $\mathcal{C} = \{\{s_0\}, \{s_1, s_2\}\}$. Clearly, $\bigcap \mathcal{C} = \emptyset$, so $f(\bigcap \mathcal{C}) = \emptyset$. However, $f(\{s_0\}) = \{t_0\}$ and $f(\{s_1, s_2\}) = \{t_0, t_1\}$, which shows that $\bigcap\{f(L) \mid L \in \mathcal{C}\} = \{t_0\}$.

**Theorem 1.67.** *Let $f : S \longrightarrow T$ be a function and let $U$ and $V$ be two subsets of $T$. Then, $f^{-1}(U - V) = f^{-1}(U) - f^{-1}(V)$.*

*Proof.* Let $s \in f^{-1}(U - V)$. We have $f(s) \in U - V$, so $f(s) \in U$ and $f(s) \notin V$. This implies $s \in f^{-1}(U)$ and $s \notin f^{-1}(V)$, so $s \in f^{-1}(U) - f^{-1}(V)$, which yields the inclusion

$$f^{-1}(U - V) \subseteq f^{-1}(U) - f^{-1}(V).$$

Conversely, let $s \in f^{-1}(U) - f^{-1}(V)$. We have $s \in f^{-1}(U)$ and $s \notin f^{-1}(V)$ which amount to $f(s) \in U$ and $f(s) \notin V$, respectively. Therefore, $f(s) \in U - V$, which implies $s \in f^{-1}(U - V)$. This proves the inclusion:

$$f^{-1}(U) - f^{-1}(V) \subseteq f^{-1}(U - V),$$

which concludes the argument. □

**Corollary 1.68.** *Let $f : S \longrightarrow T$ be a function and let $V$ be a subset of $T$. We have $f^{-1}(\bar{V}) = \overline{f^{-1}(V)}$.*

*Proof.* Note that $f^{-1}(T) = S$ for any function $f : S \longrightarrow T$. Therefore, by choosing $U = T$ in the equality of Theorem 1.67, we have

$$S - f^{-1}(V) = f^{-1}(T - V),$$

which is precisely the statement of this corollary. □

### 1.3.4 Finite and Infinite Sets

Functions allow us to compare sizes of sets. This idea is formalized next.

**Definition 1.69.** *Two sets $S$ and $T$ are* equinumerous *if there is a bijection $f : S \longrightarrow T$.*

The notion of equinumerous sets allows us to introduce formally the notions of finite and infinite sets.

**Definition 1.70.** *A set $S$ is* finite *if there exists a natural number $n \in \mathbb{N}$ such that $S$ is equinumerous with the set $\{0, \ldots, n-1\}$. Otherwise, the set $S$ is said to be* infinite.

If $S$ is an infinite set and $T$ is a subset of $S$ such that $S - T$ is finite, then we refer to $T$ as a *cofinite set*.

**Theorem 1.71.** *If $n \in \mathbb{N}$ and $f : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, n-1\}$ is an injection, then $f$ is also a surjection.*

*Proof.* Let $f : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, n-1\}$ be an injection. Suppose that $f$ is not a surjection, that is, there is $k$ such that $0 \leq k \leq n-1$ and $k \notin \text{Ran}(f)$. Since $f$ is injective, the elements $f(0), f(1), f(n-1)$ are distinct; this leads to a contradiction because $k$ is not one of them. Thus, $f$ is a surjection. □

**Theorem 1.72.** *For any natural numbers $m, n \in \mathbb{N}$, the following statements hold:*

(i) *There exists an injection from $\{0, \ldots, n-1\}$ to $\{0, \ldots, m-1\}$ if and only if $n \leq m$.*

(ii) *There exists a surjection from $\{0, \ldots, n-1\}$ to $\{0, \ldots, m-1\}$ if and only if $n \leq m > 0$ or if $n = m = 0$.*

(iii) *There exists a bijection between $\{0, \ldots, n-1\}$ and $\{0, \ldots, m-1\}$ if and only if $n = m$.*

*Proof.* For the first part of the theorem, if $n \leq m$, then the mapping $f : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, m-1\}$ given by $f(k) = k$ is the desired injection. Conversely, if $f : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, m-1\}$ is an injection, the list $(f(0), \ldots, f(n-1))$ consists of $n$ distinct elements and is a subset of the set $\{0, \ldots, m-1\}$. Therefore, $n \leq m$.

For the second part, if $n = m = 0$, then the empty function is a surjection from $\{0, \ldots, n-1\}$ to $\{0, \ldots, m-1\}$. If $n \geq m > 0$, then we can define a surjection $f : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, m-1\}$ by defining

$$f(r) = \begin{cases} r & \text{if } 0 \leq r \leq m-1, \\ 0 & \text{if } m \leq r \leq n-1. \end{cases}$$

Conversely, suppose that $f : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, m-1\}$ is a surjection. Define $g : \{0, \ldots, m-1\} \longrightarrow \{0, \ldots, n-1\}$ by defining $g(r)$, for $0 \leq r \leq m-1$, to be the least $t$, $0 \leq t \leq n-1$, for which $f(t) = r$. (Such $t$ exists since $f$ is a surjection.) Then, $g$ is an injection, and hence, by the first part, $m \leq n$. In addition, if $m = 0$, then we must also have $n = 0$ or else the function $f$ could not exist.

For Part (iii), if $n = m$, then the identity function is the desired bijection. Conversely, if there is a bijection from $\{0, \ldots, n-1\}$ to $\{0, \ldots, m-1\}$, then by the first part, $n \leq m$, while by the second part, $n \geq m$, so $n = m$. □

**Corollary 1.73.** *If $S$ is a finite set, then there is a unique natural number $n$ for which there exists a bijection from $\{0, \ldots, n-1\}$ to $S$.*

*Proof.* Suppose that $f : \{0, \ldots, n-1\} \longrightarrow S$ and $g : \{0, \ldots, m-1\} \longrightarrow S$ are both bijections. Then, $g^{-1}f : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, m-1\}$ is a bijection, so $n = m$. □

If $S$ is a finite set, we denote by $|S|$ the unique natural number that exists for $S$ according to Corollary 1.73. We refer to $|S|$ as the *cardinality* of $S$.

**Corollary 1.74.** *Let $S$ and $T$ be finite sets.*
(i) *There is an injection from $S$ to $T$ if and only if $|S| \leq |T|$.*
(ii) *There is a surjection from $S$ to $T$ if and only if $|S| \geq |T|$.*
(iii) *There is a bijection from $S$ to $B$ if and only if $|S| = |T|$.*

*Proof.* Let $|S| = n$ and $|T| = m$ and let $f : \{0, \ldots, n-1\} \longrightarrow S$ and $g : \{0, \ldots, m-1\} \longrightarrow T$ be bijections. If $h : S \longrightarrow T$ is an injection, then $g^{-1}hf : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, m-1\}$ is an injection, so by Theorem 1.72, Part (i), $n \leq m$, i.e., $|S| \leq |T|$. Conversely, if $n \leq m$, then there is an injection $k : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, m-1\}$, namely the inclusion, and $gkf^{-1} : S \longrightarrow T$ is an injection.

The other parts are proven similarly. □

**Corollary 1.75.** *Let $S$ and $T$ be two finite sets with the same cardinality. If $h : S \longrightarrow T$, then the following are equivalent:*
(i) *$h$ is an injection,*
(ii) *$h$ is a surjection, and*
(iii) *$h$ is a bijection.*

*Proof.* Let $|S| = |T| = n$ and let $f : \{0, \ldots, n-1\} \longrightarrow S$ and $g : \{0, \ldots, m-1\} \longrightarrow T$ be bijections. If $h$ is an injection, then $k = g^{-1}hf : \{0, \ldots, n-1\} \longrightarrow \{0, \ldots, n-1\}$ is also an injection, so, by Theorem 1.72, $k$ is a surjection. But then $h = gkf^{-1}$ is also a surjection.

If $h$ is a surjection, then, for each $b \in T$, $h^{-1}(\{b\})$ is nonempty. Thus, since $f$ is a surjection, for each $b \in T$, there is some $i$, $0 \leq i \leq n-1$, with $h(f(i)) = b$. We may thus define $k : T \longrightarrow S$ by defining $k(b)$ to be $f(i)$, where $i$ is the least number with $h(f(i)) = b$. Then, $k$ is a right inverse for $h$, and $k$ is an injection. By the first part of the proof, $k$ is a surjection and hence a bijection, and $h$, being a left inverse for $k$, must be $k^{-1}$, so $h$ is also a bijection.

Finally, if $h$ is a bijection, then $h$ is an injection. □

### 1.3.5 Generalized Set Products and Sequences

The Cartesian product of two sets was introduced as the set of ordered pairs of elements of these sets. Here we present a definition of an equivalent notion that can be generalized to an arbitrary family of sets.

**Definition 1.76.** *Let $S$ and $T$ be two sets. The* set product *of $S$ and $T$ is the set of functions of the form $p : \{0, 1\} \longrightarrow S \cup T$ such that $f(0) \in S$ and $f(1) \in T$.*

Note that the function $\Phi : P \longrightarrow S \times T$ given by $\Phi(p) = (p(0), p(1))$ is a bijection between the set product $P$ of the sets $S$ and $T$ and the Cartesian product $S \times T$. Thus, we can regard a function $p$ in the set product of $S$ and $T$ as an alternate representation of an ordered pair.

**Definition 1.77.** *Let $\mathcal{C} = \{S_i \mid i \in I\}$ be a collection of sets indexed by a set $I$. The* set product of $\mathcal{C}$ *is the set $\prod \mathcal{C}$ of all functions $f : I \longrightarrow \bigcup \mathcal{C}$ such that $f(i) \in S_i$ for every $i \in I$.*

*Example 1.78.* Let $\mathcal{C} = \{\{0, \ldots, i\} \mid i \in \mathbb{N}\}$ be a family of sets indexed by the set of natural numbers. Clearly, we have $\bigcup \mathcal{C} = \mathbb{N}$. The set $\prod \mathcal{C}$ consists of those functions $f$ such that $f(i) \in \{0, \ldots, i\}$ for $i \in \mathbb{N}$, that is, of those functions such that $f(i) \leq i$ for every $i \in I$.

**Definition 1.79.** *Let $\mathcal{C} = \{S_i \mid i \in I\}$ be a collection of sets indexed by a set $I$ and let $i$ be an element of $I$. The $i^{\text{th}}$* projection *is the function $p_i : \prod \mathcal{C} \longrightarrow S_i$ defined by $p_i(f) = f(i)$ for every $f \in \prod \mathcal{C}$.*

**Theorem 1.80.** *Let $\mathcal{C} = \{S_i \mid i \in I\}$ be a collection of sets indexed by a set $I$ and let $T$ be a set such that, for every $i \in I$ there exists a function $g_i : T \longrightarrow S_i$. Then, there exists a unique function $h : T \longrightarrow \prod \mathcal{C}$ such that $g_i = p_i h$ for every $i \in I$.*

*Proof.* For $t \in T$, define $h(t) = f$, where $f(i) = g_i(t)$ for every $i \in I$. We have $p_i(h(t)) = p_i(f) = g_i(t)$ for every $t \in T$, so $h$ is a function that satisfies the conditions of the statement.

Suppose now that $h_1$ is another function, $h_1 : T \longrightarrow \prod \mathcal{C}$, such that $g_i = p_i h_1$ and $h_1(t) = f_1$. We have $g_i(t) = p_i(h_1(t)) = p_i(f_1) = p_i(f)$, so $f(i) = f_1(i)$ for every $i \in I$. Thus, $f = f_1$ and $h(t) = h_1(t)$ for every $t \in T$, which shows that $h$ is unique with the property of the statement. □

**Definition 1.81.** *Let $\mathcal{C} = \{S_0, \ldots, S_{n-1}\}$ be a collection of $n$ sets indexed by the set $\{0, \ldots, n-1\}$.*

*The* set product $\prod \mathcal{P}$ *consists of those functions $f : \{0, \ldots, n-1\} \longrightarrow \bigcup_{i=0}^{n-1} S_i$ such that $f(i) \in S_i$ for $0 \leq i \leq n-1$.*

*For set products of this type, we use the alternative notation $S_0 \times \cdots \times S_{n-1}$.*

*If $S_0 = \cdots = S_{n-1} = S$, we denote the set product $S_0 \times \cdots \times S_{n-1}$ by $S^n$.*

*A* sequence on $S$ *of length $n$ is a member of this set product. If the set $S$ is clear from the context, then we refer to $\boldsymbol{s}$ as a* sequence.

*The set of finite sequences of length $n$ on the set $S$ is denoted by $\mathbf{Seq}_n(S)$.*

If $\mathbf{s} \in \mathbf{Seq}_n(S)$, we refer to the number $n$ as the *length of the sequence* $\mathbf{s}$ and it is denoted by $|\mathbf{s}|$. The *set of finite sequences on a set $S$* is the set $\bigcup\{\mathbf{Seq}_n(S) \mid n \in \mathbb{N}\}$, which is denoted by $\mathbf{Seq}(S)$.

For a sequence $\mathbf{s}$ of length $n$ on the set $S$ such that $\mathbf{s}(i) = s_i$ for $0 \leq i \leq n-1$, we denote $\mathbf{s}$ as

$$\mathbf{s} = (s_0, s_1, \ldots, s_{n-1}).$$

The elements $s_0, \ldots, s_{n-1}$ are referred to as the *components* of $\mathbf{s}$.

For a sequence $\mathbf{r} \in \mathbf{Seq}(S)$, we denote the set of elements of $S$ that occur in $\mathbf{s}$ by *set*$(\mathbf{r})$.

In certain contexts, such as the study of formal languages, sequences over a nonempty, finite set $I$ are referred to as *words*. The set $I$ itself is called an *alphabet*. We use special notation for words. If $I = \{a_0, \ldots, a_{n-1}\}$ is an alphabet and $\mathbf{s} = (a_{i_0}, a_{i_1}, \ldots, a_{i_{p-1}})$ is a word over the alphabet $I$, then we write $\mathbf{s} = a_{i_0} a_{i_1} \cdots a_{i_{p-1}}$.

The notion of a relation can also be generalized.

**Definition 1.82.** *Let $\mathcal{C} = \{C_i \mid i \in I\}$ be a collection of sets. A* $\mathcal{C}$-relation *is a subset $\rho$ of the generalized Cartesian product $\prod \mathcal{C}$. If $I$ is a finite set and $|I| = n$, then we say that $\rho$ is an* $n$-ary relation.

*For small values of $n$, we use specific terms such as* binary relation *for $n = 2$ or* ternary relation *for $n = 3$.*

*The number $n$ is the* arity of the relation $\rho$.

*Example 1.83.* Let $I = \{0, 1, 2\}$ and $C_0 = C_1 = C_2 = \mathbb{R}$. Define the ternary relation $\rho$ on the collection $\{C_0, C_1, C_2\}$ by

$$\rho = \{(x, y, z) \in \mathbb{R}^3 \mid x < y < z\}.$$

In other words, we have $(x, y, z) \in \rho$ if and only if $y \in (x, z)$.

**Definition 1.84.** *Let $\boldsymbol{p}$ and $\boldsymbol{q}$ be two finite sequences in $\mathbf{Seq}(S)$ such that $|\boldsymbol{p}| = m$ and $|\boldsymbol{q}| = n$. The* concatenation *or the* product *of $\boldsymbol{p}$ and $\boldsymbol{q}$ is the sequence $\boldsymbol{r}$ given by*

$$\boldsymbol{r}(i) = \begin{cases} \boldsymbol{p}(i) & \text{if } 0 \leq i \leq m-1 \\ \boldsymbol{q}(i-m) & \text{if } m \leq i \leq m+n-1. \end{cases}$$

*The concatenation of $\boldsymbol{p}$ and $\boldsymbol{q}$ is denoted by $\boldsymbol{pq}$.*

*Example 1.85.* Let $S = \{0, 1\}$ and let $\mathbf{p}$ and $\mathbf{q}$ be the sequences

$$\mathbf{p} = (0, 1, 0, 0, 1, 1),$$
$$\mathbf{q} = (1, 1, 1, 0).$$

By Definition 1.84, we have

$$\mathbf{pq} = (0, 1, 0, 0, 1, 1, 1, 1, 1, 0),$$
$$\mathbf{qp} = (1, 1, 1, 0, 0, 1, 0, 0, 1, 1).$$

The example above shows that, in general, $\mathbf{pq} \neq \mathbf{qp}$.

It follows immediately from Definition 1.84 that

$$\boldsymbol{\lambda}\mathbf{p} = \mathbf{p}\boldsymbol{\lambda} = \mathbf{p}$$

for every sequence $\mathbf{p} \in \mathbf{Seq}(S)$.

**Definition 1.86.** *Let $\boldsymbol{x}$ be a sequence, $\boldsymbol{x} \in \mathbf{Seq}(S)$. A sequence $\boldsymbol{y} \in \mathbf{Seq}(S)$ is:*

(i) *a* prefix *of $\boldsymbol{x}$ if $\boldsymbol{x} = \boldsymbol{yv}$ for some $\boldsymbol{v} \in \mathbf{Seq}(S)$;*
(ii) *a* suffix *of $\boldsymbol{x}$ if $\boldsymbol{x} = \boldsymbol{uy}$ for some $\boldsymbol{v} \in \mathbf{Seq}(S)$; and*
(iii) *an* infix *of $\boldsymbol{x}$ if $\boldsymbol{x} = \boldsymbol{uyv}$ for some $\boldsymbol{u}, \boldsymbol{v} \in \mathbf{Seq}(S)$.*

*A sequence $\boldsymbol{y}$ is a* proper prefix *(a* proper suffix*, a* proper infix*) of $\boldsymbol{x}$ if $\boldsymbol{y}$ is a prefix (suffix, infix) and $\boldsymbol{y} \notin \{\boldsymbol{\lambda}, \boldsymbol{x}\}$.*

*Example 1.87.* Let $S = \{a, b, c, d\}$ and $\mathbf{x} = (b, a, b, a, c, a)$. The sequence $\mathbf{y} = (b, a, b, a)$ is a prefix of $\mathbf{x}$, $\mathbf{z} = (a, c, a)$ is a suffix of $\mathbf{x}$, and $\mathbf{t} = (b, a)$ is an infix of the same sequence.

For a sequence $\mathbf{x} = (x_0, \ldots, x_{n-1})$, we denote by $\mathbf{x}_{ij}$ the infix $(x_i, \ldots, x_j)$ for $0 \leq i \leq j \leq n-1$. If $j < i$, $\mathbf{x}_{i,j} = \boldsymbol{\lambda}$.

**Definition 1.88.** *Let $S$ be a set and let $\boldsymbol{r}, \boldsymbol{s} \in \mathbf{Seq}(S)$ such that $|\boldsymbol{r}| \leq |\boldsymbol{s}|$. The sequence $\boldsymbol{r}$ is a* subsequence *of $\boldsymbol{s}$, denoted $\boldsymbol{r} \sqsubseteq \boldsymbol{s}$, if there is a function $f : \{0, \ldots, m-1\} \longrightarrow \{0, \ldots, n-1\}$ such that $f(0) < f(1) < \cdots < f(m-1)$ and $\boldsymbol{r} = \boldsymbol{s}f$.*

Note that the mapping $f$ mentioned above is necessarily injective.

If $\mathbf{r} \sqsubseteq \mathbf{s}$, as in Definition 1.88, we have $r_i = s_{f(i)}$ for $0 \leq i \leq m-1$. In other words, we can write $\mathbf{r} = (s_{i_0}, \ldots, s_{i_{m-1}})$, where $i_p = f(p)$ for $0 \leq p \leq m-1$.

The set of subsequences of a sequence $\mathbf{s}$ is denoted by $SUBSEQ(\mathbf{s})$. There is only one subsequence of $\mathbf{s}$ of length 0, namely $\boldsymbol{\lambda}$.

*Example 1.89.* For $S = \{a, b, c, d\}$ and $\mathbf{x} = (b, a, b, a, c, a)$ we have $\mathbf{y} = (b, b, c) \sqsubseteq \mathbf{x}$ because $\mathbf{y} = \mathbf{x}f$, where $f : \{0, 1, 2\} \longrightarrow \{0, 1, 2, 3, 4\}$ is defined by $f(0) = 0, f(1) = 2$, and $f(2) = 4$. Note that $set(\mathbf{y}) = \{b, c\} \subseteq set(\mathbf{x}) = \{a, b, c\}$.

**Definition 1.90.** *Let $T$ be a set. An* infinite sequence *on $T$ is a function of the form $\boldsymbol{s} : \mathbb{N} \longrightarrow T$.*

*The set of infinite sequences on $T$ is denoted by $\mathbf{Seq}_\infty(T)$. If $\boldsymbol{s} \in \mathbf{Seq}_\infty(T)$, we write $|\boldsymbol{s}| = \infty$.*

For $\mathbf{s} \in \mathbf{Seq}_\infty(T)$ such that $\mathbf{s}(n) = s_n$ for $n \in \mathbb{N}$, we also use the notation $\mathbf{s} = (s_0, \ldots, s_n, \ldots)$.

The notion of a subsequence for infinite sequences has a definition that is similar to the case of finite sequences. Let $\mathbf{s} \in \mathbf{Seq}_\infty(T)$ and let $\mathbf{r} : D \longrightarrow T$ be a function, where $D$ is either a set of the form $\{0, \ldots, m-1\}$ or the set $\mathbb{N}$. Then, $\mathbf{r}$ is a subsequence of $\mathbf{s}$ if there exists a function $f : D \longrightarrow \mathbb{N}$ such that $f(0) < f(1) < \cdots < f(k-1) < \cdots$ such that $\mathbf{r} = \mathbf{s}f$. In other words, a subsequence of an infinite sequence can be a finite sequence (when $D$ is finite) or an infinite sequence. Observe that $\mathbf{r}(k) = \mathbf{s}(f(k)) = s_{f(k)}$ for $k \in D$. Thus, as was the case for finite sequences, the members of the sequence $\mathbf{r}$ are

extracted among the members of the sequence $\mathbf{s}$. We denote this by $\mathbf{r} \sqsubseteq \mathbf{s}$, as we did for the similar notion for finite sequences.

*Example 1.91.* Let $\mathbf{s} \in \mathbf{Seq}_\infty(\mathbb{R})$ be the sequence defined by $\mathbf{s}(n) = (-1)^n$ for $n \in \mathbb{N}$, $\mathbf{s} = (1, -1, 1, -1, \ldots)$. If $f : \mathbb{N} \longrightarrow \mathbb{N}$ is the function given by $f(n) = 2n$ for $n \in \mathbb{N}$, then $\mathbf{r} = \mathbf{s}f$ is defined by $r_k = \mathbf{r}(k) = \mathbf{s}(f(k)) = (-1)^{2k} = 1$ for $k \in \mathbb{N}$.

### Occurrences in Sequences

Let $\mathbf{x}, \mathbf{y} \in \mathbf{Seq}(S)$. An *occurrence* of $\mathbf{y}$ in $\mathbf{x}$ is a pair $(\mathbf{y}, i)$ such that $0 \leq i \leq |\mathbf{x}| - |\mathbf{y}|$ and $\mathbf{y}(k) = \mathbf{x}(i + k)$ for every $k$, $0 \leq k \leq |\mathbf{y}| - 1$.

The set of all occurrences of $\mathbf{y}$ in $\mathbf{x}$ is denoted by $\mathsf{OCC}_{\mathbf{y}}(\mathbf{x})$.

There is an occurrence $(\mathbf{y}, i)$ of $\mathbf{y}$ in $\mathbf{x}$ if and only if $\mathbf{y}$ is an infix of $\mathbf{x}$. If $|\mathbf{y}| = 1$, then an occurrence of $\mathbf{y}$ in $\mathbf{x}$ is called an *occurrence of the symbol* $\mathbf{y}(0)$ in $\mathbf{x}$.

$|\mathsf{OCC}_{(s)}(\mathbf{x})|$ will be referred to as the number of occurrences of a symbol $s$ in a finite sequence $\mathbf{x}$ and be denoted by $|\mathbf{x}|_s$.

Observe that there are $|\mathbf{x}| + 1$ occurrences of the null sequence $\boldsymbol{\lambda}$ in any sequence $\mathbf{x}$.

Let $\mathbf{x} \in \mathbf{Seq}(S)$ and let $(\mathbf{y}, i)$ and $(\mathbf{y}', j)$ be occurrences of $\mathbf{y}$ and $\mathbf{y}'$ in $\mathbf{x}$. The occurrence $(\mathbf{y}', j)$ is a *part of the occurrence* $(\mathbf{y}, i)$ if $0 \leq j - i \leq |\mathbf{y}| - |\mathbf{y}'|$.

*Example 1.92.* Let $S = \{a, b, c\}$ and let $\mathbf{x} \in \mathbf{Seq}(S)$ be defined by $\mathbf{x} = (a, a, b, a, b, a, c)$. The occurrences $((a, b), 1), ((b, a), 2)$, and $((a, b), 3)$ are parts of the occurrence $((a, b, a, b), 1)$.

**Theorem 1.93.** *If* $(\boldsymbol{y}, j) \in \mathsf{OCC}_{\boldsymbol{y}}(\boldsymbol{x})$ *and* $(\boldsymbol{z}, i) \in \mathsf{OCC}_{\boldsymbol{z}}(\boldsymbol{y})$, *then* $(\boldsymbol{z}, i + j) \in \mathsf{OCC}_{\boldsymbol{z}}(\boldsymbol{x})$.

*Proof.* The argument is left to the reader. □

**Definition 1.94.** *Let* $\boldsymbol{x}$ *be a finite sequence and let* $(\boldsymbol{y}, i)$ *be an occurrence of* $\boldsymbol{y}$ *in* $\boldsymbol{x}$. *If* $\boldsymbol{x} = \boldsymbol{x}_0\boldsymbol{y}\boldsymbol{x}_1$, *where* $|\boldsymbol{x}_0| = i$, *then the sequence which results from the* replacement *of the occurrence* $(\boldsymbol{y}, i)$ *in* $\boldsymbol{x}$ *by the finite sequence* $\boldsymbol{y}'$ *is the sequence* $\boldsymbol{x}_0\boldsymbol{y}'\boldsymbol{x}_1$, *denoted by* $\mathtt{replace}\,(\boldsymbol{x}, (\boldsymbol{y}, i), \boldsymbol{y}')$.

*Example 1.95.* For the occurrences $((a, b), 1), ((a, b), 3)$ of the sequence $(a, b)$ in the sequence $\mathbf{x} = (a, a, b, a, b, a, c)$, we have

$$\mathtt{replace}\,(\mathbf{x}, ((a, b), 1), (c, a, c)) = (a, c, a, c, a, b, a, c)$$
$$\mathtt{replace}\,(\mathbf{x}, ((a, b), 3), (c, a, c)) = (a, a, b, c, a, c, a, c).$$

### Sequences of Sets

Next we examine sets defined by sequences of sets.

Let $\mathbf{s}$ be a sequence of sets. The intersection of $\mathbf{s}$ is denoted by $\bigcap_{i=0}^{n-1} S_i$ if $\mathbf{s}$ is a sequence of length $n$ and by $\bigcap_{i=0}^{\infty} S_i$ if $\mathbf{s}$ is an infinite sequence. Similarly, the union of $\mathbf{s}$ is denoted by $\bigcup_{i=0}^{n-1} S_i$ if $\mathbf{s}$ is a sequence of length $n$ and by $\bigcup_{i=0}^{\infty} S_i$ if $\mathbf{s}$ is an infinite sequence.

**Definition 1.96.** *A sequence of sets $\boldsymbol{s} = (S_0, S_1, \ldots)$ is* expanding *if $i < j$ implies $S_i \subseteq S_j$ for every $i, j$ in the domain of $s$.*

*If $i < j$ implies $S_j \subseteq S_i$ for every $i, j$ in the domain of $\boldsymbol{s}$, then we say that $\boldsymbol{s}$ is a* contracting *sequence of sets.*

*A sequence of sets is* monotonic *if it is expanding or contracting.*

**Definition 1.97.** *Let $\boldsymbol{s}$ be an infinite sequence of subsets of a set $S$, where $\boldsymbol{s}(i) = S_i$ for $i \in \mathbb{N}$.*

*The set $\bigcup_{i=0}^{\infty} \bigcap_{j=i}^{\infty} S_j$ is referred to as the* lower limit *of $\boldsymbol{s}$; the set $\bigcap_{i=0}^{\infty} \bigcup_{j=i}^{\infty} S_j$ is the* upper limit *of $\boldsymbol{s}$. These two sets will be denoted by* $\liminf \boldsymbol{s}$ *and* $\limsup \boldsymbol{s}$*, respectively.*

If $x \in \liminf \mathbf{s}$, then there exists $i$ such that $x \in \bigcap_{j=i}^{\infty} S_j$; in other words, $x$ belongs to almost all sets $S_i$.

If $x \in \limsup \mathbf{s}$, then for every $i$ there exists $j \geq i$ such that $x \in S_j$; in this case, $x$ belongs to infinitely many sets of the sequence.

Clearly, we have $\liminf \mathbf{s} \subseteq \limsup \mathbf{s}$.

**Definition 1.98.** *A sequence of sets $\boldsymbol{s}$ is* convergent *if $\liminf \boldsymbol{s} = \limsup \boldsymbol{s}$. In this case, the set $L = \liminf \boldsymbol{s} = \limsup \boldsymbol{s}$ is said to be the* limit *of the sequence $\boldsymbol{s}$.*

*The limit of $\boldsymbol{s}$ will be denoted by* $\lim \boldsymbol{s}$.

*Example 1.99.* Every expanding sequence of sets is convergent. Indeed, since $\mathbf{s}$ is expanding, we have $\bigcap_{j=i}^{\infty} S_j = S_i$. Therefore, $\liminf \mathbf{s} = \bigcup_{i=0}^{\infty} S_i$. On the other hand, $\bigcup_{j=i}^{\infty} S_j \subseteq \bigcup_{i=0}^{\infty} S_i$ and therefore $\limsup \mathbf{s} \subseteq \liminf \mathbf{s}$. This shows that $\liminf \mathbf{s} = \limsup \mathbf{s}$, that is, $\mathbf{s}$ is convergent.

A similar argument can be used to show that $\mathbf{s}$ is convergent when $\mathbf{s}$ is contracting.

Let $\mathcal{C}$ be a collection of subsets of a set $S$. Denote by $\mathcal{C}_\sigma$ the collection of all unions of subcollections of $\mathcal{C}$ indexed by $\mathbb{N}$ and by $\mathcal{C}_\delta$ the collection of all intersections of such subcollections of $\mathcal{C}$,

$$\mathcal{C}_\sigma = \left\{ \bigcup_{n \leq 0} C_n \mid C_n \in \mathcal{C} \right\},$$

$$\mathcal{C}_\delta = \left\{ \bigcap_{n \leq 0} C_n \mid C_n \in \mathcal{C} \right\}.$$

Observe that by taking $C_n = C \in \mathcal{C}$ for $n \geq 0$, it follows that $\mathcal{C} \subseteq \mathcal{C}_\sigma$ and $\mathcal{C} \subseteq \mathcal{C}_\delta$.

**Theorem 1.100.** *For any collection of subsets $\mathcal{C}$ of a set $S$, we have $(\mathcal{C}_\sigma)_\sigma = \mathcal{C}_\sigma$ and $(\mathcal{C}_\delta)_\delta = \mathcal{C}_\delta$.*

*Proof.* The argument is left to the reader. □

The operations $\sigma$ and $\delta$ can be applied iteratively. We shall denote sequences of applications of these operations by subscripts adorning the affected collection. The order of application coincides with the order of these symbols in the subscript. For example, $(\mathcal{C})_{\sigma\delta\sigma}$ means $((\mathcal{C}_\sigma)_\delta)_\sigma$. Thus, Theorem 1.100 can be restated as the equalities $\mathcal{C}_{\sigma\sigma} = \mathcal{C}_\sigma$ and $\mathcal{C}_{\delta\delta} = \mathcal{C}_\delta$.

Observe that if $\mathbf{c} = (C_0, C_1, \ldots)$ is a sequence of sets, then $\limsup \mathbf{c} = \bigcap_{i=0}^{\infty} \bigcup_{j=i}^{\infty} C_j \in \mathcal{C}_{\sigma\delta}$ and $\liminf \mathbf{c} = \bigcup_{i=0}^{\infty} \bigcap_{j=i}^{\infty} C_j$ belongs to $\mathcal{C}_{\delta\sigma}$, where $\mathcal{C} = \{C_n \mid n \in \mathbb{N}\}$.

### 1.3.6 Equivalence Relations

Equivalence relations occur in many data mining problems and are closely related to the notion of partition, which we discuss in Section 1.3.7.

**Definition 1.101.** *An* equivalence relation *on a set $S$ is a relation that is reflexive, symmetric, and transitive.*

The set of equivalences on $A$ is denoted by $EQS(S)$.

An important example of an equivalence relation is presented next.

**Definition 1.102.** *Let $U$ and $V$ be two sets, and consider a function $f : U \longrightarrow V$. The relation $\boldsymbol{ker}(f) \subseteq U \times U$, called the* kernel *of $f$, is given by*

$$\boldsymbol{ker}(f) = \{(u, u') \in U \times U \mid f(u) = f(u')\}.$$

In other words, $(u, u') \in \mathbf{ker}(f)$ if $f$ maps both $u$ and $u'$ into *the same* element of $V$.

It is easy to verify that the relation introduced above is an equivalence. Indeed, it is clear that $(u, u) \in \mathbf{ker}(f)$ for any $u \in U$, which shows that $\iota_U \subseteq \mathbf{ker}(f)$.

The relation $\mathbf{ker}(f)$ is symmetric since $(u, u') \in \mathbf{ker}(f)$ means that $f(u) = f(u')$; hence, $f(u') = f(u)$, which implies $(u', u) \in \mathbf{ker}(f)$.

Suppose that $(u, u'), (u', u'') \in \mathbf{ker}(f)$. Then, we have $f(u) = f(u')$ and $f(u') = f(u'')$, which gives $f(u) = f(u'')$. This shows that $(u, u'') \in \mathbf{ker}(f)$; hence, $\mathbf{ker}(f)$ is transitive.

*Example 1.103.* Let $m \in \mathbb{N}$ be a positive natural number. Define the function $f_m : \mathbb{Z} \longrightarrow \mathbb{N}$ by $f_m(n) = r$ if $r$ is the remainder of the division of $n$ by $m$. The range of the function $f_m$ is the set $\{0, \ldots, m-1\}$.

The relation $\mathbf{ker}(f_m)$ is usually denoted by $\equiv_m$. We have $(p, q) \in \equiv_m$ if and only if $p - q$ is divisible by $m$; if $(p, q) \in \equiv_m$, we also write $p \equiv q (\mathrm{mod}\ m)$.

**Definition 1.104.** *Let $\rho$ be an equivalence on a set $U$ and let $u \in U$.*
*The* equivalence class *of $u$ is the set $[u]_\rho$, given by*

$$[u]_\rho = \{y \in U \mid (u, y) \in \rho\}.$$

When there is no risk of confusion, we write simply $[u]$ instead of $[u]_\rho$.

Note that an equivalence class $[u]$ of an element $u$ is never empty since $u \in [u]$ because of the reflexivity of $\rho$.

**Theorem 1.105.** *Let $\rho$ be an equivalence on a set $U$ and let $u, v \in U$. The following three statements are equivalent:*
(i) $(u, v) \in \rho$;
(ii) $[u] = [v]$;
(iii) $[u] \cap [v] \neq \emptyset$.

*Proof.* The argument is immediate and we omit it. □

**Definition 1.106.** *Let $S$ be a set and let $\rho \in EQS(S)$. A subset $U$ of $S$ is* $\rho$-saturated *if it equals a union of equivalence classes of $\rho$.*

It is easy to see that $U$ is a $\rho$-saturated set if and only if $x \in U$ and $(x, y) \in \rho$ imply $y \in U$. It is clear that both $\emptyset$ and $S$ are $\rho$-saturated sets.

The following statement is immediate.

**Theorem 1.107.** *Let $S$ be a set, $\rho \in EQS(S)$, and $\mathcal{C} = \{U_i \mid i \in I\}$ be a collection of $\rho$-saturated sets. Then, both $\bigcup \mathcal{C}$ and $\bigcap \mathcal{C}$ are $\rho$-saturated sets. Also, the complement of every $\rho$-saturated set is a $\rho$-saturated set.*

*Proof.* We leave the argument to the reader. □

A more general class of relations that generalizes equivalence relations is introduced next.

**Definition 1.108.** *A* tolerance relation *(or, for short, a* tolerance *on a set $S$ is a relation that is reflexive and symmetric.*

The set of tolerances on $A$ is denoted by $TOL(S)$.

*Example 1.109.* Let $a$ be a nonnegative number and let $\rho_a \subseteq \mathbb{R} \times \mathbb{R}$ be the relation defined by

$$\rho_a = \{(x, y) \in S \times S \mid |x - y| \leq a\}.$$

It is clear that $\rho_a$ is reflexive and symmetric; however, $\rho_a$ is not transitive in general. For example, we have $(3, 5) \in \rho_2$ and $(5, 6) \in \rho_2$, but $(3, 6) \notin \rho_2$. Thus, $\rho_2$ is a tolerance but is not an equivalence.

### 1.3.7 Partitions and Covers

Next, we introduce the notion of partition of a set, a special collection of subsets of a set.

**Definition 1.110.** *Let $S$ be a nonempty set. A* partition *of $S$ is a nonempty collection of nonempty subsets of $S$, $\pi = \{B_i \mid i \in I\}$, such that $\bigcup\{B_i \mid i \in I\} = S$, and $B_i \cap B_j = \emptyset$ for every $i, j \in I$ such that $i \neq j$.*

*Each set $B_i$ of $\pi$ is a* block of the partition $\pi$.

*The set of partitions of a set $S$ is denoted by $PART(S)$. The partition of $S$ that consists of all singletons of the form $\{s\}$ with $s \in S$ will be denoted by $\alpha_S$; the partition that consists of the set $S$ itself will be denoted by $\omega_S$.*

*Example 1.111.* For the two-element set $S = \{a, b\}$, there are two partitions: the partition $\alpha_S = \{\{a\}, \{b\}\}$ and the partition $\omega_S = \{\{a, b\}\}$.

For the one-element set $T = \{c\}$, there exists only one partition, $\alpha_T = \omega_T = \{\{t\}\}$.

*Example 1.112.* A complete list of partitions of a set $S = \{a, b, c\}$ consists of the following:

$$\begin{aligned}
\pi_0 &= \{\{a\}, \{b\}, \{c\}\},\\
\pi_1 &= \{\{a, b\}, \{c\}\},\\
\pi_2 &= \{\{a\}, \{b, c\}\},\\
\pi_3 &= \{\{a, c\}, \{b\}\},\\
\pi_4 &= \{\{a, b, c\}\}.
\end{aligned}$$

Clearly, $\pi_0 = \alpha_S$ and $\pi_4 = \omega_S$.

**Definition 1.113.** *Let $S$ be a set and let $\pi, \sigma \in PART(S)$. The partition $\pi$ is* finer *than the partition $\sigma$ if every block $C$ of $\sigma$ is a union of blocks of $\pi$. This is denoted by $\pi \leq \sigma$.*

**Theorem 1.114.** *Let $\pi = \{B_i \mid i \in I\}$ and $\sigma = \{C_j \mid j \in J\}$ be two partitions of a set $S$.*

*For $\pi, \sigma \in PART(S)$, we have $\pi \leq \sigma$ if and only if for every block $B_i \in \pi$ there exists a block $C_j \in \sigma$ such that $B_i \subseteq C_j$.*

*Proof.* If $\pi \leq \sigma$, then it is clear for every block $B_i \in \pi$ there exists a block $C_j \in \sigma$ such that $B_i \subseteq C_j$.

Conversely, suppose that for every block $B_i \in \pi$ there exists a block $C_j \in \sigma$ such that $B_i \subseteq C_j$. Since two distinct blocks of $\sigma$ are disjoint, it follows that for any block $B_i$ of $\pi$ the block $C_j$ of $\sigma$ that contains $B_i$ is unique. Therefore, if a block $B$ of $\pi$ intersects a block $C$ of $\sigma$, then $B \subseteq C$.

Let $Q = \bigcup\{B_i \in \pi \mid B_i \subseteq C_j\}$. Clearly, $Q \subseteq C_j$. Suppose that there exists $u \in C_j - Q$. Then, there is a block $B_\ell \in \pi$ such that $x \in B_\ell \cap C_j$, which

implies that $B_\ell \subseteq C_j$. This means that $u \in B_\ell \subseteq C$, which contradicts the assumption we made about $x$. Consequently, $C_j = Q$, which concludes the argument. □

Note that $\alpha_S \leq \pi \leq \omega_S$ for every $\pi \in PART(S)$.

We saw that two equivalence classes either coincide or are disjoint. Therefore, starting from an equivalence $\rho \in EQS(U)$, we can build a partition of the set $U$.

**Definition 1.115.** *The* quotient set *of the set $U$ with respect to the equivalence $\rho$ is the partition $U/\rho$, where*

$$U/\rho = \{[u]_\rho \mid u \in U\}.$$

An alternative notation for the partition $U/\rho$ is $\pi_\rho$.

Moreover, we can prove that any partition defines an equivalence.

**Theorem 1.116.** *Let $\pi = \{B_i \mid i \in I\}$ be a partition of the set $U$. Define the relation $\rho_\pi$ by $(x, y) \in \rho_\pi$ if there is a set $B_i \in \pi$ such that $\{x, y\} \subseteq B_i$. The relation $\rho_\pi$ is an equivalence.*

*Proof.* Let $B_i$ be the block of the partition that contains $u$. Since $\{u\} \subseteq B_i$, we have $(u, u) \in \rho_\pi$ for any $u \in U$, which shows that $\rho_\pi$ is reflexive.

The relation $\rho_\pi$ is clearly symmetric. To prove the transitivity of $\rho_\pi$, consider $(u, v), (v, w) \in \rho_\pi$. We have the blocks $B_i$ and $B_j$ such that $\{u, v\} \subseteq B_i$ and $\{v, w\} \subseteq B_j$. Since $v \in B_i \cap B_j$, we obtain $B_i = B_j$ by the definition of partitions; hence, $(u, w) \in \rho_\pi$. □

**Corollary 1.117.** *For any equivalence $\rho \in EQS(U)$, we have $\rho = \rho_{\pi_\rho}$. For any partition $\pi \in PART(U)$, we have $\pi = \pi_{\rho_\pi}$.*

*Proof.* The argument is left to the reader. □

The previous corollary amounts to the fact that there is a bijection $\phi : EQS(U) \longrightarrow PART(U)$, where $\phi(\rho) = \pi_\rho$. The inverse of this mapping, $\Psi : PART(U) \longrightarrow EQS(U)$, is given by $\psi(\pi) = \rho_\pi$.

Also, note that, for $\pi, \pi' \in PART(S)$, we have $\pi \leq \pi'$ if and only if $\rho_\pi \subseteq \rho_{\pi'}$.

We say that a subset $T$ of a set $S$ is *$\pi$-saturated* if it is a $\rho_\pi$-saturated set.

**Theorem 1.118.** *For any mapping $f : U \longrightarrow V$, there is a bijection $h : U/\mathbf{ker}(f) \longrightarrow f(U)$.*

*Proof.* Consider the $\mathbf{ker}(f)$ class $[u]$ of an element $u \in U$, and define $h([x]) = f(x)$. The mapping $h$ is well-defined for if $u' \in [u]$, then $(u, u') \in \mathbf{ker}(f)$, which gives $f(u) = f(u')$.

Further, $h$ is onto since if $y \in f(U)$, then there is $u \in U$ such that $f(u) = y$, and this gives $y = h([u])$.

To prove the injectivity of $h$, assume that $h([u]) = h([v])$. This means that $f(u) = f(v)$; hence, $(u, v) \in \mathbf{ker}(f)$, which means, of course, that $[u] = [v]$. □

An important consequence of the previous proposition is the following decomposition theorem for mappings.

**Theorem 1.119.** *Every mapping $f : U \longrightarrow V$ can be decomposed as a composition of three mappings: a surjection $g : U \longrightarrow U/\mathbf{ker}(f)$, a bijection $h : U/\mathbf{ker}(f) \longrightarrow f(U)$, and an injection $k : f(U) \longrightarrow V$.*

*Proof.* The mapping $g : U \longrightarrow U/\mathbf{ker}(f)$ is defined by $g(u) = [u]$ for $u \in U$, while $k : f(A) \longrightarrow B$ is the inclusion mapping given by $k(v) = v$ for all $v \in f(U)$. Therefore, $k(h(g(u))) = k(h([u])) = k(f(u)) = f(u)$ for all $u \in U$. □

A generalization of the notion of partition is introduced next.

**Definition 1.120.** *Let $S$ be a set. A* cover *of $S$ is a nonempty collection $\mathcal{C}$ of nonempty subsets of $S$, $\mathcal{C} = \{B_i \mid i \in I\}$, such that $\bigcup\{B_i \mid i \in I\} = S$.*

*The set of covers of a set $S$ is denoted by $COVERS(S)$.*

*Example 1.121.* Let $S$ be a set. The collection $\mathcal{P}_k(S)$ of subsets of $S$ that contain $k$ elements is a cover of $S$ for every $k \geq 1$. For $k = 1$, $\mathcal{P}_1(S)$ is actually the partition $\alpha_S$.

The notion of collection refinement introduced in Definition 1.12 is clearly applicable to covers and will be used in Section 12.4.

## 1.4 The Axiom of Choice

The Axiom of Choice, one of the fundamental principles of set theory was formulated by Ernst Zermello at the beginning of the twentieth century. To state this axiom, we need the notion introduced next.

**Definition 1.122.** *Let $\mathcal{C}$ be a nonempty collection of nonempty sets that are pairwise disjoint. A* selective set *for $\mathcal{C}$ is a set $K$ such that for every set $S \in \mathcal{C}$ the intersection $K \cap M$ contains exactly one element.*

A related notion is the notion of a choice function.

**Definition 1.123.** *A* choice function *for a collection of sets $\mathcal{C}$ is a mapping $f : \mathcal{C} \longrightarrow \bigcup \mathcal{C}$, such that, for each $S \in \mathcal{C}$ if $S \neq \emptyset$, then $f(S) \in S$.*

*A* selective function *for set $M$ is a choice function for the collection $\mathcal{P}(M)$.*

Attempts at proving this axiom from other principles of set theory have failed; instead many equivalent formulations of the Axiom of Choice were obtained, and we present a few of these formulations in this section.

**Axiom of Choice** Every collection of sets has a choice function.

Note that the Axiom of Choice merely states the existence of a choice function but states no criterion for choosing its values.

**Theorem 1.124.** *The Axiom of Choice is equivalent to the following statements:*
(i) *Every collection of nonempty, pairwise disjoint sets has a selective set.*
(ii) *For every set, there exists a selective function.*
(iii) *The Cartesian product of every family of nonempty sets is nonempty.*

*Proof.* `The Axiom of Choice implies (i):` Consider a collection $\mathcal{C}$ that consists of nonempty sets that are pairwise disjoint. The Axiom of Choice implies the existence of a selective function $f$ for $\mathcal{C}$.

We claim that the set $K = \{f(M)|M \in \mathcal{C}\}$ is a selective set for $\mathcal{C}$. Indeed, we have $f(M) \in K \cap M$ for every $M \in \mathcal{C}$, and since the sets of $\mathcal{C}$ are pairwise disjoint, it follows that $K \cap M$ consists of exactly one element $f(M)$.

(i) implies (ii): If $M = \emptyset$, then the function $f : \mathcal{P}(\emptyset) \longrightarrow \{\emptyset\}$ defined by $f(\emptyset) = \emptyset$ can serve as a selective function. Therefore, we may assume that $M$ is nonempty. Consider the collection $\mathcal{C}$ of all sets having the form $K \times \{K\}$ for $K \subseteq M$. Observe that, if $K \neq H$, the sets $K \times \{K\}$ and $H \times \{H\}$ are disjoint. Indeed, if we have $t \in (K \times \{K\}) \cap (H \times \{H\})$, then $t = (x, K) = (y, H)$, and this implies $K = H$.

Let $D$ be a selective set for $\mathcal{C}$. For every $K \subseteq M$, there is $x \in K$ such that $D \cap (K \times \{K\}) = \{(x, K)\}$. We can define the selective function on $M$ by $f(K) = x$, where $(x, K) \in D$. Since $x \in K$, it follows that $f$ is indeed a selective function for the set $M$.

(ii) implies (iii): Consider a collection of nonempty sets $\mathcal{C} = \{M_i|i \in I\}$, and let $M = \bigcup\{M_i|i \in I\}$. If $f$ is a selective function for $M$, then for every $M_i \in \mathcal{C}$, we have $f(M_i) \in M_i$. Consider the mapping $t : I \longrightarrow M$ defined by $t(i) = f(M_i)$. We have $t \in \prod_{i \in I} M_i$ by Definition 1.77.

(iii) implies the Axiom of Choice: Let $\mathcal{C} = \{M_i|i \in I\}$ be a collection of sets. Consider the set of indices $J = \{i|i \in I, M_i \in \mathcal{C}, M_i \neq \emptyset\}$ and the collection $\mathcal{D} = \{M_i|M_i, i \in J\}$.

Let $t \in \prod \mathcal{D}$. For every $i \in J$, we have $t(i) \in M_i$, which means that for every $M_i \in \mathcal{C}$, $M_i \neq \emptyset$, we have $t(i) \in M_i$. This means that $t$ is a choice function for $\mathcal{C}$. □

## 1.5 Countable Sets

A set is called *countable* if it is either empty or the range of a sequence. A set that is not countable is called *uncountable*.

Note that if $S$ is a countable set and $f : S \longrightarrow T$ is a surjection, then $T$ is also countable.

*Example 1.125.* Every finite set is countable. Let $S$ be a finite set. If $S = \emptyset$, then $S$ is countable. Otherwise, suppose that $S = \{a_0, \ldots, a_{n-1}\}$, where $n \geq 1$. Define the sequence $\mathbf{s}$ as

$$\mathbf{s}(i) = \begin{cases} a_i & \text{if } 0 \leq i \leq n-1, \\ a_{n-1} & \text{otherwise.} \end{cases}$$

It is immediate that $\text{Ran}(\mathbf{s}) = S$.

*Example 1.126.* The set $\mathbb{N}$ is countable because $\mathbb{N} = \text{Ran}\,\mathbf{s}$, where $\mathbf{s}$ is the sequence $\mathbf{s}(n) = n$ for $n \in \mathbb{N}$. A similar argument can be used to show that the set $\mathbb{Z}$ is countable. Indeed, let $\mathbf{t}$ be the sequence defined by

$$\mathbf{t}(n) = \begin{cases} \frac{n-1}{2} & \text{if } n \text{ is odd} \\ -\frac{n}{2} & \text{if } n \text{ is even.} \end{cases}$$

Let $m$ be an integer. If $m > 0$, then $m = \mathbf{t}(2m-1)$; otherwise (that is, if $m \leq 0$), $m = \mathbf{t}(-2m)$, so $\mathbf{z} = \text{Ran}(\mathbf{t})$.

*Example 1.127.* We shall prove now that the set $N \times N$ is countable. To this end, consider the representation of pairs of natural numbers shown in Figure 1.2. The pairs of the set $\mathbb{N} \times \mathbb{N}$ are scanned in the order suggested by the dotted arrows. The $0^{\text{th}}$ pair is $(0,0)$, followed by $(0,1)$, $(1,0)$, $(0,2)$, $(1,1)$, $(2,0)$, etc. We define the bijection $\beta : \mathbb{N} \times \mathbb{N} \longrightarrow \mathbb{N}$ as $\beta(p,q) = n$, where $n$ is the place occupied by the pair $(p,q)$ in the previous list. Thus, $\beta(0,0) = 0$, $\beta(0,1) = 1$, $\beta(2,0) = 5$, and so on.

In general, bijections of the form $h : \mathbb{N} \times \mathbb{N} \longrightarrow \mathbb{N}$ are referred to as *pairing functions*, so $\beta$ is an example of a pairing function.

The existence of the inverse bijection $\beta^{-1} : \mathbb{N} \longrightarrow \mathbb{N} \times \mathbb{N}$ shows that $\mathbb{N} \times \mathbb{N}$ is indeed a countable set because $\mathbb{N} \times \mathbb{N} = \text{Ran}(\beta^{-1})$.

Another example of a bijection between $\mathbb{N} \times \mathbb{N}$ and $\mathbb{P}$ can be found in Exercise 22.

Starting from countable sets, it is possible to construct uncountable sets, as we see in the next example.

*Example 1.128.* Let $F$ be the set of all functions of the form $f : \mathbb{N} \longrightarrow \{0,1\}$. We claim that $F$ is not countable.

If $F$ were countable, we could write $F = \{f_0, f_1, \ldots, f_n, \ldots\}$. Define the function $g : \mathbb{N} \longrightarrow \{0,1\}$ by $g(n) = \overline{f_n(n)}$ for $n \in \mathbb{N}$, where $\overline{0} = 1$ and $\overline{1} = 0$. Note that $g \neq f_n$ for every $f_n$ in $F$ because $g(n) = \overline{f_n(n)} \neq f_n(n)$, that is, $g$ is different from $f_n$ at least on $n$ for every $n \in \mathbb{N}$. But $g$ is a function defined on $\mathbb{N}$ with values in $\{0,1\}$, so it must equal some function $f_m$ from $F$. This contradiction implies that $F$ is not countable.

**Theorem 1.129.** *A subset $T$ of a countable set $S$ is countable.*

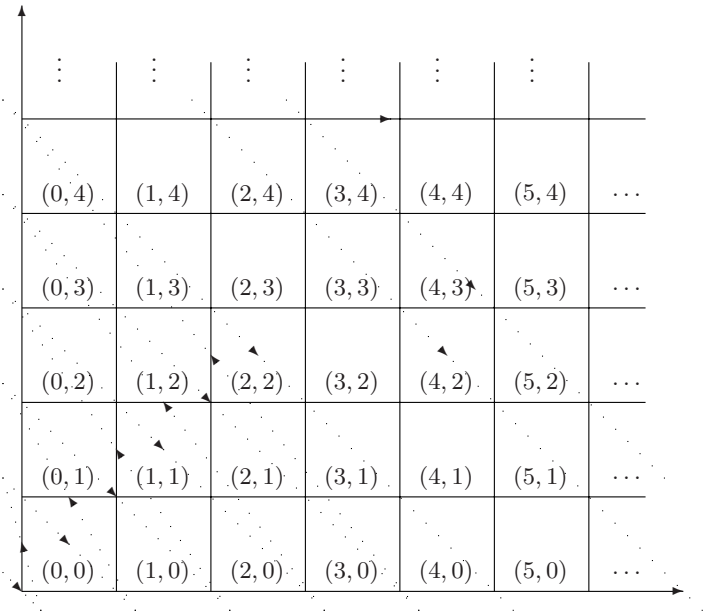



**Fig. 1.2.** Representation of $\mathbb{N} \times \mathbb{N}$.

*Proof.* If either $S$ or $T$ are empty, the statement is immediate. Suppose therefore that neither $S$ nor $T$ are empty and that $S = \{a_0, \ldots, a_{n-1}\}$. Since $T$ is a subset of $S$, we can write $T = \{a_{i_0}, \ldots, a_{i_{m-1}}\}$, so $T$ is the range of the sequence $\mathbf{t} : \{0, \ldots, m-1\} \longrightarrow S$ given by $\mathbf{t}(j) = a_{i_j}$ for $0 \leq j \leq m-1$. Thus, $T$ is countable. □

**Theorem 1.130.** *Let $\mathcal{C} = \{C_i \mid i \in I\}$ be a collection of sets such that each set $C_i$ is countable and the indexing set is countable. Then, $\bigcup \mathcal{C}$ is a countable set.*

*Proof.* Without loss of generality, we can assume that none of the sets $C_i$ is empty. Also, if $I = \emptyset$, then $\bigcup \mathcal{C}$ is empty and therefore countable.

Suppose, therefore, that $C_i$ is the range of the sequence $\mathbf{s}_i$ for $i \in I$ and that $I \neq \emptyset$. Since $I$ is a countable set, we can assume that $I$ is the range of a sequence $\mathbf{z}$.

Define the function $f : \mathbb{N} \times \mathbb{N} \longrightarrow \bigcup \mathcal{C}$ by $f(p, q) = s_{z(p)}(q)$ for $p, q \in \mathbb{N}$. It is easy to verify that $f$ is a surjection. Indeed, if $c \in \bigcup \mathcal{C}$, there exists a set $C_i$ such that $c \in C_i$ and, since $C_i$ is the range of the sequence $\mathbf{s}_i$, it follows that $c = \mathbf{s}_i(q)$ for some $q \in \mathbb{N}$.

Suppose that $i = \mathbf{z}(p)$. Then, we can write $c = \mathbf{s}_{\mathbf{z}(p)}(q) = f(p, q)$, which allows us to conclude that $\bigcup \mathcal{C} = \text{Ran}(f)$. To obtain the enumerability of

$\bigcup \mathcal{C}$, observe that this set can now be regarded as the range of the sequence $\mathbf{u}$ given by $\mathbf{u}(n) = f(\beta^{-1}(n))$, where $\beta$ is the pairing function introduced in Example 1.127. □

**Theorem 1.131.** *The set $\mathbb{Q}$ of rational numbers is countable.*

*Proof.* We show first that the set of positive rational numbers $\mathbb{Q}_{\geq 0}$ is countable. Indeed, note that the function $f : \mathbb{N} \times \mathbb{N} \longrightarrow \mathbb{Q}_{\geq 0}$ defined by $f(m, n) = \frac{m}{n+1}$ is a surjection. A similar argument shows that the set of negative rational numbers is countable. By Theorem 1.130, the countability of $\mathbb{Q}$ follows immediately. □

## 1.6 Elementary Combinatorics

In this section, we discuss some counting techniques for certain collections of objects. We begin with a study of bijections of finite sets that is useful for the presentation of these techniques.

**Definition 1.132.** *A* permutation *of a set $S$ is a bijection $f : S \longrightarrow S$.*

A permutation $f$ of a finite set $S = \{s_0, \ldots, s_{n-1}\}$ is completely described by the sequence $(f(s_0), \ldots, f(s_{n-1}))$. No two distinct components of such a sequence may be equal because of the injectivity of $f$, and all elements of the set $S$ appear in this sequence because $f$ is surjective. Therefore, the number of permutations equals the number of such sequences, which allows us to conclude that there are $n(n-1)\cdots 2 \cdot 1$ permutations of a finite set $S$ with $|S| = n$.

The number $n(n-1)\cdots 2 \cdot 1$ is usually denoted by $n!$. This notation is extended by defining $0! = 1$, which is consistent with the interpretation of $n!$ as the number of bijections of a set that has $n$ elements.

The *set of permutations* of the set $S = \{1, \ldots, n\}$ is denoted by $PERM_n$. If $f \in PERM_n$ is such a permutation, we write

$$f : \begin{pmatrix} 1 & \cdots & i & \cdots & n \\ a_1 & \cdots & a_i & \cdots & a_n \end{pmatrix},$$

where $a_i = f(i)$ for $1 \leq i \leq n$. To simplify the notation, we shall specify $f$ just by the sequence $(a_1, \ldots, a_i, \ldots, a_n)$.

**Definition 1.133.** *Let $S$ be a finite set, $f$ be a permutation of $S$, and $x \in S$. The* cycle of $x$ *is the set of elements of the form $C_{f,x} = \{f^i(x) \mid i \in \mathbb{N}\}$. The number $|C_{f,x}|$ is the* length of the cycle.

*Cycles of length 1 are said to be* trivial.

Let $S$ be a finite set. Since $C_{f,x} \subseteq S$, it is clear that $C_{f,x}$ is a finite set. If $|C_{f,x}| = \ell$, then

$$C_{f,x} = \{x, f(x), \ldots, f^{\ell-1}(x)\}.$$

Note that each pair of elements $f^i(x)$ and $f^j(x)$ are distinct for $0 \leq i, j \leq \ell - 1$, and $i \neq j$ because otherwise we would have $|C_{f,x}| < \ell$. Moreover, $f^\ell(x) = x$.

If $z \in C_{f,x}$, then $z = f^k(x)$ for some $k$, $0 \leq k \leq \ell - 1$, where $\ell = |C_{f,x}|$. Since $x = f^\ell(x)$, it follows that $x = f^{\ell-k}(z)$, which shows that $x \in C_{f,z}$. Thus, $C_{f,x} = C_{f,z}$.

Thus, the cycles of a permutation of a finite set $S$ form a partition $\pi_f$ of $S$.

**Definition 1.134.** *A* $k$-cyclic permutation *of a finite set $S$ is a permutation such that $\pi_f$ consists of a cycle of length $k$ and a number of $|S| - k$ cycles of length* 1.

*A* transposition *of $S$ is a* 2*-cyclic permutation.*

Note that if $f$ is a transposition of $S$, then $f^2 = 1_S$.

**Theorem 1.135.** *Let $S$ be a finite set, $f$ be a permutation, and $\pi_f = \{C_{f,x_1}, \ldots, C_{f,x_m}\}$ be the cycle partition associated to $f$. Define the cyclic permutations $g_1, \ldots, g_m$ of $S$ as*

$$g_p(t) = \begin{cases} f(t) & \text{if } t \in C_{f,x_p}, \\ t & \text{otherwise.} \end{cases}$$

*Then, $g_p g_q = g_q g_p$ for every $p, q$ such that $1 \leq p, q \leq m$.*

*Proof.* Observe first that $u \in C_{f,x}$ if and only if $f(x) \in C_{f,x}$ for any cycle $C_{f,x}$.

We can assume that $p \neq q$. Then, the cycles $C_{f,x_p}$ and $C_{f,x_q}$ are disjoint. If $u \notin C_{f,x_p} \cup C_{f,x_q}$, then we can write $g_p(g_q(u)) = g_p(u) = u$ and $g_q(g_p(u)) = g_q(u) = u$.

Suppose now that $u \in C_{f,x_p} - C_{f,x_q}$. We have $g_p(g_q(u)) = g_p(u) = f(u)$. On the other hand, $g_q(g_p(u)) = g_q(f(u)) = f(u)$ because $f(u) \notin C_{f,x_q}$. Thus, $g_p(g_q(u)) = g_q(g_p(u))$. The case where $u \in C_{f,x_q} - C_{f,x_p}$ is treated similarly. Also, note that $C_{f,x_p} \cap C_{f,x_q} = \emptyset$, so, in all cases, we have $g_p(g_q(x)) = g_q(g_p(u))$. □

The set of cycles $\{g_1, \ldots, g_m\}$ is the *cyclic decomposition of the permutation $f$.*

**Definition 1.136.** *A* standard transposition *is a transposition that changes the places of two adjacent elements.*

*Example 1.137.* The permutation $f \in PERM_5$ given by

$$f : \begin{pmatrix} 1\ 2\ 3\ 4\ 5 \\ 1\ 3\ 2\ 4\ 5 \end{pmatrix}$$

is a standard transposition of the set $\{1, 2, 3, 4, 5\}$.

On the other hand, the permutation

$$g : \begin{pmatrix} 1\ 2\ 3\ 4\ 5 \\ 1\ 5\ 3\ 4\ 2 \end{pmatrix}$$

is a transposition but not a standard transposition of the same set because the pair of elements involved is not consecutive.

If $f \in PERM_n$ is specified by the sequence $(a_1, \ldots, a_n)$, we refer to each pair $(a_i, a_j)$ such that $i < j$ and $a_i > a_j$ as an *inversion* of the permutation $f$. The set of all such inversions will be denoted by $INV(f)$. The number of elements of $INV(f)$ is denoted by $inv(f)$.

A *descent* of a permutation $f$ of $S = \{1, \ldots, n\}$ is a number $j$ such that $1 \leq j \leq n-1$ and $a_j > a_{j+1}$. The set of descents of $f$ is denoted by $D(f)$.

*Example 1.138.* Let $f \in PERM_6$ be:

$$f : \begin{pmatrix} 1\ 2\ 3\ 4\ 5\ 6 \\ 4\ 2\ 5\ 1\ 6\ 3 \end{pmatrix}.$$

We have

$$INV(f) = \{(4,2), (4,1), (4,3), (2,1), (5,1), (5,3), (6,3)\}.$$

and $inv(f) = 7$. Furthermore, $D(f) = \{1, 3, 5\}$.

It is easy to see that the following conditions are equivalent for a permutation $f$ of the finite set $S$:

(i) $f = 1_S$;
(ii) $inv(f) = 0$;
(iii) $D(f) = \emptyset$.

**Theorem 1.139.** *Every permutation $f \in PERM_n$ can be written as a composition of transpositions.*

*Proof.* If $D(f) = \emptyset$, then $f = 1_S$ and the statement is vacuous. Suppose therefore that $D(f) \neq \emptyset$, and let $j \in D(f)$, which means that $(a_j, a_{j+1})$ is an inversion $f$. Let $g$ be the standard transposition that exchanges $a_j$ and $a_{j+1}$. It is clear that $inv(gf) = inv(f) - 1$. Thus, if $g_i$ are the transpositions that correspond to all standard inversions of $f$ for $1 \leq i \leq p = inv(f)$, it follows that $g_p \cdots g_1 f$ has 0 inversions and, as observed above, $g_p \cdots g_1 f = 1_S$. Since $g^2 = 1_S$ for every transposition $g$, we have $f = g_p \cdots g_1$, which gives the desired conclusion. □

**Theorem 1.140.** *If $f$ is a permutation of the finite set $S$, then $inv(f)$ is the least number of standard transpositions, and the number of standard transpositions involved in any other factorization of $f$ as a product of standard transposition differs from $inv(f)$ by an even number.*

*Proof.* Let $f = h_q \cdots h_1$ be a factorization of $f$ as a product of standard transpositions. Then, $h_1 \cdots h_q f = 1_S$ and we can define the sequence of permutations $f_l = h_l \cdots h_1 f$ for $1 \leq l \leq q$. Since each $h_i$ is a standard transposition, we have $inv(f_{l+1}) - \inf(f_l) = 1$ or $inv(f_{l+1}) - \inf(f_l) = -1$. If

$$|\{l \mid 1 \leq l \leq q-1 \text{ and } inv(f_{l+1}) - inv(f_l) = 1\}| = r,$$

then $|\{l \mid 1 \leq l \leq q-1 \text{ and } inv(f_{l+1}) - inv(f_l) = -1\}| = q - r$, so $inv(f) + r - (q-r) = 0$, which means that $q = inv(f) + 2r$. This implies the desired conclusion. □

**Definition 1.141.** *A permutation $f$ of $\{1, \ldots, n\}$ is even* (odd) if $inv(f)$ is an even (odd) number.

Note that any transposition is an odd permutation.

**Theorem 1.142.** *The set of subsets of a set that contains $n$ elements consists of $2^n$ subsets.*

*Proof.* Let $S$ be a set that contains $n$ elements. The argument is by induction on $n$.

In the basis step, $n = 0$, so $S$ is the empty subset and $\mathcal{P}(S)$ is nonempty because it contains $\emptyset$; thus, $|\mathcal{P}(\emptyset)| = 1$.

Suppose now that $S$ contains $n$ elements, say $S = \{s_0, \ldots, s_{n-1}\}$, and let $S' = S - \{s_{n-1}\}$. Any subset $Z$ of $S$ belongs to one of the following two disjoint classes:

1. $s_{n-1} \notin Z$, so $Z \subseteq S'$, or
2. $s_{n-1} \in Z$, so $Z - s_{n-1} \subseteq S'$.

By the inductive hypothesis, both these collections contain $2^{n-1}$ subsets. Therefore, $\mathcal{P}(S)$ contains $2 \cdot 2^{n-1} = 2^n$ subsets. □

Let $S$ be a finite nonempty set, $S = \{s_1, \ldots, s_n\}$. We seek to count the sequences of $S$ having length $k$ without repetitions.

Suppose initially that $k \geq 1$. For the first place in a sequence $\mathbf{s}$ of length $k$, we have $n$ choices. Once an element of $S$ has been chosen for the first place, since the sequence may not contain repetitions, we have $n-1$ choices for the second place, etc. For the $k^{\text{th}}$ component of $\mathbf{s}$, there are $n - 1 + k$ choices. Thus, the number of sequences of length $k$ without repetitions is given by $n(n-1)\cdots(n-k+1)$. We shall denote this number by $A(n,k)$.

There exists only one sequence of length 0, namely the empty sequence, so we extend the definition of $A$ by $A(n,0) = 1$ for every $n \in \mathbb{N}$.

An important special case of this counting problem occurs when $k = n$. In this case, a sequence of length $n$ without repetitions is essentially a permutation of the set $S$. Thus, the number of permutations of $S$ is $n(n-1)\cdots 1$; this number is denoted by $n!$, and we refer to it as the *factorial of* $n$.

In the case $n = 0$, we define $0! = 1$ to maintain consistency with the definition of $A(n, 0)$.

**Theorem 1.143.** *Let $S$ and $T$ be two finite sets. We have*

$$|S \cup T| = |S| + |T| - |S \cap T|,$$
$$|S \oplus T| = |S| + |T| - 2 \cdot |S \cap T|.$$

*Proof.* If $S \cap T = \emptyset$, then $S \cup T = S \oplus T$ and the equalities above are obviously true. Therefore, we may assume that $S \cap T = \{z_1, \ldots, z_p\}$, where $p \geq 1$. Thus, the sets $S$ and $T$ can be written as $S = \{x_0, \ldots, x_{m-1}, z_1, \ldots, z_p\}$ and $T = \{y_0, \ldots, y_{n-1}, z_1, \ldots, z_p\}$. The symmetric difference $S \oplus T$ can be written as

$$S \oplus T = \{x_0, \ldots, x_{m-1}, y_0, \ldots, y_{n-1}\}.$$

Since $|S| = m + p$, $T = n + p$, $|S \cup T| = m + n + p$, and $|S \oplus T| = m + p$, the equalities of the theorem follow immediately. □

Let us now count the number of $k$-element subsets of a set that contains $n$ elements.

Let $S$ be a set such that $|S| = n$. Define the equivalence $\sim$ on the set $\mathbf{Seq}(S)$ by $\mathbf{s} \sim \mathbf{t}$ if there exists a bijection $f$ such that $\mathbf{s} = \mathbf{t}f$.

It is easy to verify that $\sim$ is an equivalence, and we leave it to the reader to perform this verification. If $\mathbf{s} : \{0, \ldots, p-1\} \longrightarrow S$ and $\mathbf{t} : \{0, \ldots, q-1\} \longrightarrow S$, $\mathbf{s} \sim \mathbf{t}$, and $f : \{0, \ldots, p-1\} \longrightarrow \{0, \ldots, q-1\}$ is a bijection, then we have $p = q$, by Theorem 1.72.

If $T$ is a subset of $S$ such that $|T| = k$, there exists a bijection $\mathbf{t} : \{0, \ldots, k-1\} \longrightarrow T$; clearly, this is a sequence without repetitions and there exist $A(n, k)$ such sequences. Note that if $\mathbf{u}$ is an equivalent sequence (that is, if $\mathbf{t} \sim \mathbf{u}$), then the range of this sequence is again the set $T$ and there are $k!$ such sequences (due to the existence of the $k!$ permutations $f$) that correspond to the same set $T$. Therefore, we may conclude that $\mathcal{P}_k(S)$ contains $\frac{A(n,k)}{k!}$ elements. We denote this number by $\binom{n}{k}$ and we refer to it as the $(n, k)$-*binomial coefficient*. We can write $\binom{n}{k}$ using factorials:

$$\begin{aligned}\binom{n}{k} &= \frac{A(n,k)}{k!} = \frac{n(n-1)\cdots(n-k+1)}{k!} \\ &= \frac{n(n-1)\cdots(n-k+1)(n-k)\cdots 2 \cdot 1}{k!(n-k)!} \\ &= \frac{n!}{k!(n-k)!}.\end{aligned}$$

We mention the following useful identities:

$$k\binom{n}{k} = n\binom{n-1}{k-1}, \tag{1.1}$$

$$\binom{n}{m} = \frac{n}{m}\binom{n-1}{m-1}. \tag{1.2}$$

Equality (1.1) can be extended as

$$k(k-1)\cdots(k-\ell)\binom{n}{k} = n(n-1)\cdots(n-\ell)\binom{n-\ell-1}{k-\ell-1} \tag{1.3}$$

for $0 \leq \ell \leq k-1$.

Consider now the $n$-degree polynomial in $x$

$$p(x) = (x+a_0)\cdots(x+a_{n-2})(x+a_{n-1}).$$

Observe that the coefficient of $x^{n-k}$ consists of the sum of all monomials of the form $a_{i_0}\cdots a_{i_{k-1}}$, where the subscripts $i_0,\ldots,i_{k-1}$ are distinct. Thus, the coefficient of $x^{n-k}$ contains $\binom{n}{k}$ terms corresponding to the $k$-element subsets of the set $\{0,\ldots,n-1\}$. Consequently, the coefficient of $x^{n-k}$ in the power $(x+a)^n$ can be obtained from the similar coefficient in $p(x)$ by taking $a_0 = \cdots = a_{n-1} = a$; thus, the coefficient is $\binom{n}{k}a^k$. This allows us to write:

$$(x+a)^n = \sum_{k=0}^{n}\binom{n}{k}x^{n-k}a^k. \tag{1.4}$$

This equality is known as *Newton's binomial formula* and has numerous applications.

*Example 1.144.* If we take $x = a = 1$ in Formula (1.4) we obtain the identity

$$2^n = \sum_{k=0}^{n}\binom{n}{k}. \tag{1.5}$$

Note that this equality can be obtained directly by observing that the right member enumerates the subsets of a set having $n$ elements by their cardinality $k$.

A similar interesting equality can be obtained by taking $x = 1$ and $a = -1$ in Formula (1.4). This yields

$$\begin{aligned} 0 &= \sum_{k=0}^{n}\binom{n}{k}(-1)^k \\ &= \binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots \\ &\quad -\binom{n}{0} - \binom{n}{2} - \binom{n}{4} - \cdots. \end{aligned}$$

This inequality shows that each set contains an equal number of subsets having an even or odd number of elements.

*Example 1.145.* Consider the equality $(x + a)^n = (x + a)^{n-1}(x + a)$. The coefficient of $x^{n-k}a^k$ in the left member is $\binom{n}{k}$. In the right member $x^{n-k}a^k$ has the coefficient $(\binom{n-1}{k} + \binom{n-1}{k-1})$, so we obtain the equality

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}, \tag{1.6}$$

for $0 \leq k \leq n-1$.

Multinomial coefficients are generalizations of binomial coefficients that can be introduced as follows. The $n^{\text{th}}$ power of the sum $x_1 + \cdots + x_k$ can be written as

$$(x_1 + \cdots + x_k)^n = \sum_{(r_1,\ldots,r_k)} c(n, r_1, \ldots, r_k) x_1^{r_1} \cdots x_k^{r_k},$$

where the sum involves all $(r_1, \ldots, r_k) \in \mathbb{N}^k$ such that $\sum_{i=1}^k r_i = n$. By analogy with the binomial coefficients, we denote $c(n, r_1, \ldots, r_k)$ by $\binom{n}{r_1,\ldots,r_n}$. As we did with binomial coefficients in Example 1.145, starting from the equality $(x_1 + \cdots + x_k)^n = (x_1 + \cdots + x_k)^{n-1}(x_1 + \cdots + x_k)$, the coefficient of the monomial $x_1^{r_1} \cdots x_k^{r_k}$ in the right member is $\binom{n}{r_1,\ldots,r_n}$. On the left member, the same coefficient is

$$\sum_{i=1}^{k} \binom{n-1}{r_1, \ldots, r_i - 1, \ldots, r_n},$$

so we obtain the identity

$$\binom{n}{r_1, \ldots, r_n} = \sum_{i=1}^{k} \binom{n-1}{r_1, \ldots, r_i - 1, \ldots, r_n}, \tag{1.7}$$

a generalization of the identity (1.6).

## 1.7 Multisets

Multisets generalize the notion of a set by allowing multiple copies of an element. Formally, we have the following definition.

**Definition 1.146.** *A* multiset *on a set $S$ is a function $M : S \longrightarrow \mathbb{N}$. Its* carrier *is the set $carr(M) = \{x \in S \mid M(x) > 0\}$. The* multiplicity *of an element $x$ of $S$ in the multiset $M$ is the number $M(x)$.*

*The set of all multisets on $S$ is denoted by $\mathcal{M}(S)$.*

*Example 1.147.* Let PRIMES be the set of prime numbers:

$$\mathsf{PRIMES} = \{2, 3, 5, 7, 11, \ldots\}. \tag{1.8}$$

A number is determined by the multiset of its prime divisors in the following sense. If $n \in \mathbb{N}$, $n \geq 1$, can be factored as a product of prime numbers, $n = p_{i_1}^{k_1} \cdots p_{i_\ell}^{k_\ell}$, where $p_i$ is the $i^{\text{th}}$ prime number and $k_1, \ldots, k_\ell$ are positive numbers, then the multiset of its prime divisors is the multiset $M_n : \mathsf{PRIMES} \longrightarrow \mathbb{N}$, where $M_n(p)$ is the exponent of the prime number $p$ in the product (1.8).

For example, $M_{1960}$ is given by

$$M_{1960}(p) = \begin{cases} 3 & \text{if } p = 2, \\ 1 & \text{if } p = 5, \\ 2 & \text{if } p = 7. \end{cases}$$

Thus, $carr(M_{1960}) = \{2, 5, 7\}$.

Note that if $m, n \in \mathbb{N}$, we have $M_m = M_n$ if and only if $m = n$.

We denote a multiset by using square brackets instead of braces. If $x$ has the multiplicity $n$ in a multiset $M$, we write $x$ a number of times $n$ inside the square brackets. For example, the multiset of Example 1.147 can be written as $[2, 2, 2, 5, 7, 7]$.

Note that while multiplicity counts in a multiset, order does not matter; therefore, the multiset $[2, 2, 2, 5, 7, 7]$ could also be denoted by $[5, 2, 7, 2, 2, 7]$ or $[7, 5, 2, 7, 2, 2]$. We also use the abbreviation $n * x$ in a multiset to mean that $x$ has the multiplicity $n$ in $M$. For example, the multiset $M_{1960}$ can be written as $M_{1960} = [3 * 2, 1 * 5, 2 * 7]$.

The multiset $M$ on the set $S$ defined by $M(x) = 0$ for $x \in S$ is the *empty multiset*.

Multisets can be combined to construct new multisets. Common set-theoretical operations such as union and intersection have natural generalizations to multisets.

**Definition 1.148.** *Let $M$ and $N$ be two multisets on a set $S$.*

*The* union of $M$ and $N$ *is the multiset $M \cup N$ defined by*

$$(M \cup N)(x) = \max\{M(x), N(x)\}$$

*for $x \in S$.*

*The* intersection of $M$ and $N$ *is the multiset $M \cap N$ defined by*

$$(M \cap N)(x) = \min\{M(x), N(x)\}$$

*for $x \in S$.*

*The* sum of $M$ and $N$ *is the multiset $M + N$ given by*

$$(M + N)(x) = M(x) + N(x)$$

*for $x \in S$.*

*Example 1.149.* Let $m, n \in \mathbb{N}$ be two numbers that have the prime factorizations

$$m = p_{i_1}^{k_1} \cdots p_{i_r}^{k_r},$$
$$n = p_{j_1}^{h_1} \cdots p_{h_s}^{h_s},$$

and let $M_m, M_n$ be the multisets of their prime divisors, as defined in Example 1.147. Denote by $\gcd(m, n)$ the greatest common divisor of $m$ and $n$, and by $\mathrm{lcm}(m, n)$ the least common multiple of these numbers.

We have

$$M_{\gcd(m,n)} = M_m \cap M_n,$$
$$M_{\mathrm{lcm}(m,n)} = M_m \cup M_n,$$
$$M_{mn} = M_m + M_n,$$

as the reader can easily verify.

A multiset on the set $\mathcal{P}(S)$ is referred to as a *multicollection* of sets on $S$.

## 1.8 Relational Databases

Relational databases are the mainstay of contemporary databases. The principles of relational databases were developed by C. D. Codd in the early 1970s [30, 31], and various extensions have been considered since. In this section, we illustrate applications of several notions introduced earlier to the formalization of database concepts.

The notion of a tabular variable (or relational variable) was introduced by C. J. Date in [35]; we also formalize the notion of table of a relational variable. To reflect the implementations of a relational database system we assume that table contents are sequences of tuples (and not just sets of tuples, a simplification often adopted in the literature that is quite distant from reality).

Let $\mathcal{U}$ be a countably infinite injective sequence having pairwise distinct members, $\mathcal{U} = (A_0, A_1, \ldots)$. The components of $\mathcal{U}$ are referred to as *attributes* and denoted, in general, by capital letters from the beginning of the alphabet, $A, B, C, \ldots$. We also consider a collection of sets indexed by the components of $\mathcal{U}$, $\mathcal{D} = \{D_A \mid A \in U\}$. The set $D_A$ is referred to as the *domain* of the attribute $A$ and denoted alternatively as $\mathrm{Dom}(A)$. We assume that each set $D_A$ contains at least two elements.

Let $H$ be a finite subset of $set(\mathcal{U})$, $H = \{A_{i_1}, \ldots, A_{i_p}\}$. We refer to such a set as a *heading*. In keeping with the tradition of the field of relational databases, we shall denote $H$ as $H = A_{i_1} \cdots A_{i_p}$. For example, instead of writing $H = \{A, B, C, D, E\}$, we shall write $H = ABCDE$.

The *set of tuples on* $H = A_{i_1} \cdots A_{i_p}$ is the set $D_{A_{i_1}} \times \cdots \times D_{A_{i_p}}$ denoted by $tupl(H)$. Thus, a *tuple* $t$ *on the heading* $H = A_{i_1} \cdots A_{i_p}$ is a sequence $t = (t_1, \ldots, t_p)$ such that $t_j \in \mathrm{Dom}(A_{i_j})$ for $1 \leq j \leq p$.

A *tabular variable* is a pair $\tau = (T, H)$, where $T$ is a word over an alphabet to be defined later and $H$ is a heading.

A value of a tabular variable $\tau = (T, H)$ is a triple $\theta = (T, H, \mathbf{r})$, where $\mathbf{r}$ is a sequence on $tupl(H)$. We refer to such a triple as a *table of the tabular variable* $\tau$ or, a $\tau$*-table*; when the tabular variable is clear from the context or irrelevant, we refer to $\theta$ just as a *table*.

The set $set(\mathbf{r})$ of tuples that constitute the components of a tuple sequence $\mathbf{r}$ is a $p$-ary relation on the collection of sets $tupl(H)$; this justifies the term "relational" used for the basic database model.

*Example 1.150.* Consider a tabular variable that is intended to capture the description of a collection of objects,

$$\tau = (\text{OBJECTS}, \text{shape length width height color}),$$

where

$$\text{Dom}(shape) = \text{Dom}(color) = \{a, \ldots, z\}^* \text{ and}$$
$$\text{Dom}(length) = \text{Dom}(width) = \text{Dom}(color) = \mathbb{N}.$$

A value of this variable is

$$(\text{OBJECTS}, \text{shape length width height color}, \mathbf{r}),$$

where $\mathbf{r}$ consists of the tuples

$$(\text{cube}, 5, 5, 5, \text{red}),$$
$$(\text{sphere}, 3, 3, 3, \text{blue}),$$
$$(\text{pyramid}, 5, 6, 4, \text{blue}),$$
$$(\text{cube}, 2, 2, 2, \text{red}),$$
$$(\text{sphere}, 3, 3, 3, \text{blue}),$$

that belong to $tupl$(shape length width height color). It is convenient to represent this table graphically as

OBJECTS

| shape | length | width | height | color |
|---|---|---|---|---|
| cube | 5 | 5 | 5 | red |
| sphere | 3 | 3 | 3 | blue |
| pyramid | 5 | 6 | 4 | blue |
| cube | 2 | 2 | 2 | red |
| sphere | 3 | 3 | 3 | blue |

The set $set(\mathbf{r})$ of tuples that corresponds to the sequence $\mathbf{r}$ of tuples of the table is

$$set(\mathbf{r}) = \{(\text{cube}, 5, 5, 5, \text{red}), (\text{sphere}, 3, 3, 3, \text{blue}),$$
$$(\text{pyramid}, 5, 6, 4, \text{blue}), (\text{cube}, 2, 2, 2, \text{red})\}.$$

Note that duplicate tuples do not exist in $set(\mathbf{r})$.

We can now formalize the notion of a relational database.

**Definition 1.151.** *A* relational database *is a finite, nonempty collection* $\mathcal{D}$ *of tabular variables* $\tau_i = (T_k, H_k)$*, where* $1 \leq k \leq m$ *such that* $i \neq j$ *implies* $T_i \neq T_j$ *for* $1 \leq i, j \leq m$*.*

In other words, a relational database $\mathcal{D}$ is a finite collection of tabular variables that have pairwise distinct names.

Let $\mathcal{D} = \{\tau_1, \ldots, \tau_m\}$ be a relational database. A *state of* $\mathcal{D}$ is a sequence of tables $(\theta_1, \ldots, \theta_m)$ such that $\theta_i$ is a table of $\tau_i$ for $1 \leq i \leq m$. The set of states of a relational database $\mathcal{D}$ will be denoted by $\mathcal{S}_{\mathcal{D}}$.

To discuss further applications we need to introduce table projection, an operation on tables that allows us to build new tables by extracting "vertical slices" of the original tables.

**Definition 1.152.** *Let* $\theta = (T, H, \boldsymbol{r})$ *be a table, where* $H = A_1 \cdots A_p$ *and* $\boldsymbol{r} = (t_1, \ldots, t_n)$*, and let* $K = A_{i_1} \cdots A_{i_q}$ *be a subsequence of set*$(H)$*.*

*The* projection *of a tuple* $t \in tupl(H)$ *on* $K$ *is the tuple* $t[K] \in tupl(K)$ *defined by* $t[K](j) = t(i_j)$ *for every* $j$*,* $1 \leq i \leq q$*.*

*The* projection *of the table* $\theta$ *on* $K$ *is the table* $\theta[K] = (T[K], K, \boldsymbol{r}[K])$*, where* $\boldsymbol{r}[K]$ *is the sequence* $(t_1[K], \ldots, t_n[K])$*.*

Observe that, for every tuple $t \in tupl(H)$, we have $t[\emptyset] = \boldsymbol{\lambda}$; also, $t[H] = t$.

*Example 1.153.* The projection of the table

OBJECTS

| shape | length | width | height | color |
|---|---|---|---|---|
| cube | 5 | 5 | 5 | red |
| sphere | 3 | 3 | 3 | blue |
| pyramid | 5 | 6 | 4 | blue |
| cube | 2 | 2 | 2 | red |
| sphere | 3 | 3 | 3 | blue |

on the set $K =$ shape color is the table

OBJECTS[shape color]

| shape | color |
|---|---|
| cube | red |
| sphere | blue |
| pyramid | blue |
| cube | red |
| sphere | blue |

Two simple but important properties of projection are given next.

**Theorem 1.154.** *Let $H$ be a set of attributes, $u, v \in tupl(H)$, and let $K$ and $L$ be two subsets of $H$. The following statements hold:*
(i) $u[K][K \cap L] = u[L][K \cap L] = u[K \cap L]$.
(ii) *The equality $u[KL] = v[KL]$ holds if and only if $u[K] = v[K]$ and $u[L] = v[L]$.*

*Proof.* The argument is a straightforward application of Definition 1.152 and is left to the reader. □

## Exercises and Supplements

1. Prove that for any set $S$ we have $\bigcup \mathcal{P}(S) = S$.
2. A set is *transitive* if $X \subseteq \mathcal{P}(X)$. Prove that $\{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}\}$ is transitive.
3. Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of sets such that $\mathcal{C} \subseteq \mathcal{D}$. Prove that $\bigcup \mathcal{C} \subseteq \bigcup \mathcal{D}$; also, if $\mathcal{C} \neq \emptyset$, then show that $\bigcap \mathcal{C} \supseteq \bigcap \mathcal{D}$.
4. Let $\{\mathcal{C}_i \mid i \in I\}$ be a family of hereditary collections of sets. Prove that $\bigcap_{i \in I} \mathcal{C}_i$ is also a hereditary collection of sets.
5. Let $\mathcal{C}$ be a nonempty collection of nonempty subsets of a set $S$. Prove that $\mathcal{C}$ is a partition of $S$ if and only if every element $a \in S$ belongs to exactly one member of the collection $\mathcal{C}$.
6. Let $S$ be a set and let $U$ be a subset of $S$. For $a \in \{0, 1\}$, define the set

$$U^a = \begin{cases} U & \text{if } a = 1, \\ S - U & \text{if } a = 0. \end{cases}$$

   a) Prove that if $\mathcal{D} = \{D_1, \ldots, D_r\}$ is a finite collection of subsets of $S$, then the nonempty sets that belong to the collection

$$\{D_1^{a_1} \cap D_2^{a_2} \cap \cdots \cap D_r^{a_r} \mid (a_1, a_2, \ldots, a_r) \in \{0, 1\}^r\},$$

   constitute a partition $\pi_{\mathcal{D}}$ of $S$.
   b) Prove that each set of $\mathcal{D}$ is a $\pi_{\mathcal{C}}$-saturated set.

   **Solution:** For $a = (a_1, \ldots, a_r) \in \{0, 1\}^r$, denote by $\mathcal{D}^{\mathbf{a}}$ the set $D_1^{a_1} \cap D_2^{a_2} \cap \cdots \cap D_r^{a_r}$.

   Let $\mathbf{a}, \mathbf{b} \in \{0, 1\}^r$ such that $\mathbf{a} \neq \mathbf{b}$ and $\mathbf{a} = (a_1, \ldots, a_r)$ and $\mathbf{b} = (b_1, \ldots, b_r)$. Note that $\mathcal{D}^{\mathbf{a}} \cap \mathcal{D}^{\mathbf{b}} = \emptyset$. Further, let $x \in S$. Define $d_i$ as $d_i = 1$ if $x \in D_i$ and $d_i = 0$ otherwise for $1 \leq i \leq r$, and let $\mathbf{d} = (d_1, \ldots, d_r)$. Then, it is clear that $x \in \mathcal{D}^{\mathbf{d}}$ and therefore $S = \bigcup \{\mathcal{D}^{\mathbf{d}} \mid \mathbf{d} \in \{0, 1\}^r\}$. This concludes the argument for the first part.

   For the second part, note that each set $D_i \in \mathcal{D}$ can be written as

$$D_i = \bigcup \{D_1^{a_1} \cap D_2^{a_2} \cap \cdots \cap D_i \cap \cdots \cap D_r^{a_r} \mid (a_1, \ldots, a_{i-1}, 1, a_{i+1}, \ldots, a_r) \in \{0, 1\}^r\}.$$

7. Prove that if $\pi, \pi' \in PART(S)$ and $\pi' \leq \pi$, then every $\pi$-saturated set is a $\pi'$-saturated set.
8. Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of subsets of a set $S$. Prove that if $T$ is a subset of $S$, then $(\mathcal{C} \cup \mathcal{D})_T = \mathcal{C}_T \cup \mathcal{D}_T$ and $(\mathcal{C} \cap \mathcal{D})_T \subseteq \mathcal{C}_T \cap \mathcal{D}_T$.
9. Let $S$ be a set and let $\mathcal{C}$ be a collection of subsets of $S$. The elements $x$ and $y$ of $S$ are *separated by* $\mathcal{C}$ if there exists $C \in \mathcal{C}$ such that either $x \in C$ and $y \notin C$ or $x \notin C$ and $y \in C$. Let $\rho \subseteq S \times \mathcal{C}$ be the relation defined in Example 1.47.
   Prove that $x$ and $y$ are separated by $\mathcal{C}$ if and only if $\rho(x) \neq \rho(y)$.
10. Prove that for all sets $R, S, T$ we have
    a) $(R \cup S) \oplus (R \cap S) = R \oplus S$,
    b) $R \cap (S \oplus T) = (R \cap S) \oplus (R \cap T)$.
11. Let $P$ and $Q$ be two subsets of a set $S$.
    a) Prove that $P \cup Q = S$ if and only if $S - P \subseteq Q$.
    b) Prove that $P \cap Q = \emptyset$ if and only if $Q \subseteq S - P$.
12. Let $S$ be a nonempty set and let $s_0$ be a fixed element of $S$. Define the collection $[x, y] = \{\{x, s_0\}, \{y, \{s_0\}\}\}$. Prove that if $[x, x'] = [y, y']$, then $x = y$ and $x' = y'$.
13. Let $S$ and $T$ be two sets. Suppose that the function $p : S \times T \longrightarrow T$ given by $p(x, y) = y$ for $x \in S$ and $y \in T$ is a bijection. What can be said about the set $S$?
14. Let $S$ and $T$ be two sets. The functions $p_1 : S \times T \longrightarrow S$ and $p_2 : S \times T \longrightarrow T$ defined by $p_1(x, y) = x$ and $p_2(x, y) = y$ for $x \in S$ and $y \in T$ are the *projections* of the Cartesian product $S \times T$ on $S$ and $T$, respectively. Let $U$ be a set such that $f : U \longrightarrow S$ and $g : U \longrightarrow T$ are two functions. Prove that there is a unique function $h : U \longrightarrow S \times T$ such that $p_1 h = f$ and $p_2 h = g$.
15. Let $\mathcal{C}$ be a collection of subsets of a set $S$. Define the relations $\rho_{\mathcal{C}}$ and $\sigma_{\mathcal{C}}$ on $S$ by

$$\sigma_{\mathcal{C}} = \{(x, y) \in S \times S \mid |x \in C \text{ if and only if } y \in C \text{ for every } C \in \mathcal{C}\}$$

    and

$$\rho_{\mathcal{C}} = \{(x, y) \in S \times S \mid |x \in C \text{ implies } y \in C \text{ for every } C \in \mathcal{C}\}.$$

    Prove that, for every collection $\mathcal{C}$, the relation $\sigma_{\mathcal{C}}$ is an equivalence and that $\rho_{\mathcal{C}}$ is a reflexive and transitive relation.
16. Prove that a relation $\rho$ is a function if and only if $\rho^{-1}\rho \subseteq \iota_{\text{Ran}(\rho)}$.
17. Prove that $\rho$ is a one-to-one relation if and only if $\rho\rho^{-1} \subseteq \iota_{\text{Dom}(\rho)}$.
18. Prove that $\rho$ is a total relation from $A$ to $B$ if and only if $\iota_A \subseteq \rho\rho^{-1}$.
19. Prove that $\rho$ is an onto relation from $A$ to $B$ if and only if $\iota_B \subseteq \rho^{-1}\rho$.
20. Prove that the composition of two injections (surjections, bijections) is an injection (a surjection, a bijection, respectively).
21. Let $f : S_1 \longrightarrow S_2$ be a function. Prove that, for every set $L \in \mathcal{P}(S_2)$, we have

$$S_1 - f^{-1}(S_2 - L) = f^{-1}(L).$$

22. Prove that the function $\gamma : \mathbb{N} \times \mathbb{N} \longrightarrow \mathbb{P}$ defined by $\gamma(p, q) = 2^p(2q + 1)$ for $p, q \in \mathbb{N}$ is a bijection.
23. Let $f : S \longrightarrow T$ be a function. Prove that $f$ is an injection if and only if $f(U \cap V) = f(U) \cap f(V)$ for every $U, V \in \mathcal{P}(S)$.
24. Let $S$ be a set and let $f_i : S \longrightarrow S$ be $m$ injective mappings for $1 \leq\leq m$. If $\mathbf{s} = (i_1, i_2, \ldots, i_k) \in \mathbf{Seq}(\{1, \ldots, m\})$ let $f_{\mathbf{s}}$ be the injection $f_{i_1} f_{i_2} \cdots f_{i_k}$ and let $T_{\mathbf{s}}$ be the set $f_{i_1}(f_{i_2}(\cdots(f_{i_k}(T)\cdots)))$.
    a) Prove that if $(i_1, i_2, \ldots) \in \mathbf{Seq}_\infty(\{1, \ldots, m\})$, then $T \supseteq T_{i_1} \supseteq T_{i_1 i_2} \supseteq \cdots$.
    b) Prove that if $\{T_{i_1}, T_{i_2}, \ldots, T_{i_m}\}$ is a partition of the set $T$, then $\{T_{\mathbf{u}} \mid \mathbf{u} \in \mathbf{Seq}_p(\{1, \ldots, n\})\}$ is a partition of $T$ for every $p \geq 1$.

    **Solution:** One can prove by induction on $|\mathbf{s}|$ that $T_{\mathbf{s}i} \subseteq T_{\mathbf{s}}$ for every sequence $\mathbf{s} \in \mathbf{Seq}(\{1, \ldots, m\})$ and $i \in \{1, \ldots, m\}$. This implies the first statement. A proof of the second statement can be obtained by induction on $p$.
25. Let $f : S \longrightarrow T$ be a function. Prove that for every $U \in \mathcal{P}(S)$ we have $U \subseteq f^{-1}(f(U))$ and for every $V \in \mathcal{P}(T)$ we have $f(f^{-1}(V)) = V$.
26. Let $S$ be a finite set and let $\mathcal{C}$ be a collection of subsets of $S$. For $x \in S$, define the mapping $\phi_x : \mathcal{C} \longrightarrow \mathcal{P}(S)$ by

$$\phi_x(C) = \begin{cases} C - \{x\} & \text{if } x \in C \text{ and } C - \{x\} \notin \mathcal{C}, \\ C & \text{otherwise,} \end{cases}$$

    for $C \in \mathcal{C}$. Prove that $|\mathcal{C}| = |\phi_x(C) \mid C \in \mathcal{C}|$.

    **Solution:** To prove the equality, it suffices to show that $\phi_x$ is injective. Suppose that $\phi_x(C_1) = \phi_x(C_2)$. Observe that in this case both $\phi_x(C_1)$ and $\phi_x(C_2)$ are computed by applying the same case of the definition of $\phi_x$. Indeed, if this were not the case, suppose that $\phi_x(C_1)$ is obtained by applying the first case and $\phi_x(C_2)$ is obtained by applying the second case. This entails $C_1 - \{x\} = C_2$, $x \in C_1$, and $C_1 - \{x\} \notin \mathcal{C}$, which is contradictory. Thus, we have $C_1 = C_2$.
27. Let $\mathcal{C} = \{C_i \mid i \in I\}$ and $\mathcal{D} = \{D_i \mid i \in I\}$ be two collections of sets indexed by the same set $I$. Define the collections

$$\mathcal{C} \vee_I \mathcal{D} = \{C_i \vee D_i \mid i \in I\},$$
$$\mathcal{C} \wedge_I \mathcal{D} = \{C_i \wedge D_i \mid i \in I\}.$$

    Prove that

$$\left(\prod \mathcal{C}\right) \cap \left(\prod \mathcal{D}\right) = \prod (\mathcal{C} \wedge_I \mathcal{D}),$$
$$\left(\prod \mathcal{C}\right) \cup \left(\prod \mathcal{D}\right) \subseteq \prod (\mathcal{C} \vee_I \mathcal{D}).$$

28. Prove that the relation $\rho \subseteq S \times S$ is

a) reflexive if $\iota_S \subseteq \rho$,
b) irreflexive if $\iota_S \cap \rho = \emptyset$,
c) symmetric if $\rho^{-1} = \rho$,
d) antisymmetric if $\rho \cap \rho^{-1} \subseteq \iota_S$,
e) asymmetric if $\rho^{-1} \cap \rho = \emptyset$,
f) transitive if $\rho^2 \subseteq \rho$.

29. Prove that if $S$ is a finite set such that $|S| = n$, then there are $2^{n^2}$ binary relations on $S$.
30. Prove that there are $2^{n(n-1)}$ binary reflexive relations on a finite set $S$ that has $n$ elements.
31. Prove that the number of antisymmetric relations on a finite set that has $n$ elements is $2^n \cdot 3^{\frac{n(n-1)}{2}}$.
32. Let $S$ be a set and let $\rho$ be a relation on $S$. Prove that $\rho$ is an equivalence on $S$ if and only if there exists a collection $\mathcal{C}$ of pairwise disjoint subsets of $S$ such that $S = \bigcup \mathcal{C}$ and $\rho = \bigcup \{C \times C \mid C \in \mathcal{C}\}$.
33. Let $\rho$ and $\rho'$ be two equivalence relations on the set $S$. Prove that $\rho \cup \rho'$ is an equivalence on $S$ if and only if $\rho\rho' \cup \rho'\rho \subseteq \rho \cup \rho'$.
34. Let $\mathcal{E} = \{\rho_i \mid i \in I\}$ be a collection of equivalence relations on a set $S$ such that, for $\rho_i, \rho_j \in \mathcal{E}$, we have $\rho_i \subseteq \rho_j$ or $\rho_j \subseteq \rho_i$, where $i, j \in I$. Prove that $\bigcup \mathcal{E}$ is an equivalence on $S$.
35. Let $\rho$ be a relation on a set $S$. Prove that the relation $\sigma = \bigcup_{n \in \mathbb{N}} (\rho \cup \rho^{-1} \cup \iota_S)^n$ is the least equivalence on $S$ that includes $\rho$.
36. Let $\mathbf{x} = (x_0, \ldots, x_{n-1})$ be a sequence in $\mathbf{Seq}(\mathbb{R})$, where $n \geq 2$. The sequence is said to be *unimodal* if there exists $j$, $0 \leq j \leq n-1$ such that $x_0 \leq x_1 \leq \cdots \leq x_j$ and $x_j \geq x_{j+1} \geq \cdots \geq x_n$.
Prove that if $\mathbf{x} \in \mathbf{Seq}(\mathbb{R}_{>0})$ and $x_{p-1}x_{p+1} \leq x_p^2$ for $1 \leq p \leq n-2$, then $\mathbf{x}$ is a unimodal sequence.
37. Let $p_1, p_2, p_3, \ldots$ be the sequence of prime numbers $2, 3, 5, \cdots$. Define the function $f : \mathbf{Seq}(\mathbb{N}) \longrightarrow \mathbb{N}$ by $f(n_1, \ldots, n_k) = p_1^{n_1} \cdots p_k^{n_k}$. Prove that $f(n_1, \ldots, n_k) = f(m_1, \ldots, m_k)$ implies $(n_1, \ldots, n_k) = (m_1, \ldots, m_k)$.
38. Prove that the Axiom of Choice is equivalent to the statement "every surjective function has a right inverse".

    **Solution:** Let $f : U \longrightarrow V$ be a surjective function. If $V = \emptyset$, then $X = \emptyset$, so $f$ is the empty function. The desired right inverse is also the empty function. Suppose that $V \neq \emptyset$ and consider the nonempty collection $\mathcal{C}_f = \{f^{-1}(v) \mid v \in V\}$ that consists of nonempty sets, that are pairwise disjoint. Let $K$ be a selective set for $\mathcal{C}_f$ and let $h : V \longrightarrow U$ be the function defined by $h(v) = u$ if $K \cap f^{-1}(y) = \{u\}$. It is easy to see that $h$ is a right inverse for $f$.

    Conversely, let $\mathcal{C}$ be a nonempty collection of nonempty, pairwise disjoint sets and let $f : \bigcup \mathcal{C} \longrightarrow \mathcal{C}$ be the function defined by $f(x) = C$ if $x \in C$. It is clear that $f$ is a surjection, so it has a right inverse, $h : \mathcal{C} \longrightarrow \bigcup \mathcal{C}$. Then, $K = h(\mathcal{C})$ is a selective set for $\mathcal{C}$.
39. Give a selective set for the collection of sets $\mathcal{C} = \{(z, z+1) \mid z \in \mathbb{Z}\}$.

40. Let $f : S \longrightarrow T$ be a function such that $f^{-1}(t)$ is a countable set for every $t \in T$. Prove that the set $S$ is countable.
41. Prove that $\mathcal{P}(\mathbb{N})$ is not countable.
42. Let $\mathbf{S} = (S_0, S_1, \ldots)$ be a sequence of countable sets. Prove that $\liminf \mathbf{S}$ and $\limsup \mathbf{S}$ are both countable sets.
43. Let $S$ be a countable set. Prove that $\mathbf{Seq}(S)$ is countable. How about $\mathbf{Seq}_\infty(S)$?
44. Let $f$ and $g$ be two transpositions on a set $S$. Prove that there is $i \in \{1, 2, 3\}$ such that $(fg)^i = 1_S$.
45. Let $x \in \mathbb{R}$ and let $m \in \mathbb{N}$. Prove that if $x \geq 0$ and $m \geq 1$, then
$$\frac{x^{m-1}}{(m-1)!} + \frac{x^m}{m!} \leq \frac{(x+1)^m}{m!}.$$
46. Starting from Newton's binomial formula, prove the following identities:
$$\sum_{k=0}^{n} (n-k)\binom{n}{k} = n2^{n-1},$$
$$\sum_{k=0}^{n} \frac{\binom{n}{k}}{n-k+1} = \frac{2^{n+1}-1}{n+1}.$$
47. Prove that
$$\binom{m+n}{k} = \sum_{i=0}^{k} \binom{m}{i}\binom{n}{k-i}$$
for $m, n, k \in \mathbb{N}$ and $k \leq m + n$.
48. Prove that $\max\{\binom{n}{k} \mid 0 \leq k \leq n\} = \binom{n}{\lfloor \frac{n}{2} \rfloor} = \binom{n}{\lceil \frac{n}{2} \rceil}$.
49. Prove that
$$\frac{2^{2n}}{2n+1} \leq \binom{2n}{n} \leq 2^{2n}$$
for $n \geq 0$.
50. Let $S$ and $T$ be two finite sets such that $|S| = m$ and $|T| = n$.
    a) Prove that the set of functions $S \longrightarrow T$ contains $n^m$ elements.
    b) Prove that the set of partial functions $S \rightsquigarrow T$ contains $(n+1)^m$ elements.
51. Let $I$ be a finite set. A *system of distinct representatives* for a collection of sets $\mathcal{C} = \{C_i \mid i \in I\}$ is an injection $r : I \to \bigcup \mathcal{C}$ such that $r(i) \in C_i$ for $i \in I$.
    Define the mapping $\Phi_{\mathcal{C}} : \mathcal{P}(I) \leq \mathcal{P}(\bigcup \mathcal{C})$ by $\Phi_{\mathcal{C}}(L) = \bigcup_{i \in L} C_i$ for $L \subseteq I$.
    a) Show that if $\mathcal{C}$ has a system of distinct representatives, then $|\Phi_{\mathcal{C}}(L)| \geq |L|$ for every $L$ such that $L \subseteq \{1, \ldots, n\}$.
    b) A subset $L$ of $I$ is *$\Phi$-critical* if $|\Phi_{\mathcal{C}}(L)| = |L|$. Let $x \in \bigcup \mathcal{C}$. Define $\Phi' : \mathcal{P}(I) \leq \mathcal{P}(\bigcup \mathcal{C})$ by $\Phi'(L) = \Phi_{\mathcal{C}}(L) - \{x\}$. Prove that if no nonempty set $L$ is $\Phi$-critical, then $|\Phi'(L)| \geq |L|$ for every $L$.

c) Let $L$ be a nonempty minimal $\Phi_{\mathcal{C}}$-critical set such that $L \subset I$. Define the collection $\mathcal{D} = \{C_i - \Phi(L) \mid i \in I - L\}$. Prove that $|\Phi_{\mathcal{D}}(H)| \geq |H|$ for every $H \subseteq I - L$.

d) Prove, by induction on the number $n = |I|$, the converse of the first statement: If $|\Phi_{\mathcal{C}}(L)| \geq |L|$ for every $L$ in $\mathcal{P}(I)$, then a system of distinct representatives exists for the collection $\mathcal{C}$ (*Hall's matching theorem*).

52. Prove the inequality

$$\binom{n}{i-1}\binom{n}{i+1} \leq \left(\binom{n}{i}\right)^2$$

for $1 \leq i \leq n-1$.

53. Let $\mathcal{C}$ be a collection of subsets of a finite set $S$.
    a) Prove that if $C \cap D \neq \emptyset$ for every pair $(C, D)$ of members of $\mathcal{C}$, then $|\mathcal{C}| \leq 2^{|S|-1}$.
    b) Prove that if $C \cup D \subset S$ for every pair $(C, D)$ of members of $\mathcal{C}$, then $|\mathcal{C}| \leq 2^{|S|-1}$.

54. Let $\mathcal{C}$ be a collection of subsets of a finite set $S$ such that $\mathcal{C} \subseteq \mathcal{P}_k(S)$ for some $k < |S|$. The *shadow* of $\mathcal{C}$ is the collection $\Delta\mathcal{C} = \{D \in \mathcal{P}_{k-1}(S) \mid D \subseteq C \text{ for some } C \in \mathcal{C}\}$. The *shade* of $\mathcal{C}$ is the collection $\nabla\mathcal{C} = \{D \in \mathcal{P}_{k+1}(S) \mid C \subseteq T \text{ for some } C \in \mathcal{C}\}$.
    Prove that:
    a) $|\Delta\mathcal{C}| \geq \frac{k}{n-k+1}|\mathcal{C}|$ for $k > 0$;
    b) $|\nabla\mathcal{C}| \geq \frac{n-k}{k+1}|\mathcal{C}|$ for $k < n$;
    c) $\frac{|\Delta\mathcal{C}|}{\binom{n}{k-1}} \geq \frac{|\mathcal{C}|}{\binom{n}{k}}$ for $k > 0$;
    d) $\frac{|\nabla\mathcal{C}|}{\binom{n}{k+1}} \geq \frac{|\mathcal{C}|}{\binom{n}{k}}$ for $k < n$;
    e) if $k \leq \frac{n-1}{2}$, then $|\nabla\mathcal{C}| \geq |\mathcal{C}|$;
    f) if $k \geq \frac{n+1}{2}$, then $|\Delta\mathcal{C}| \geq |\mathcal{C}|$.

    **Solution:** We discuss only the first inequality since the argument for the second is similar. If $C \in \mathcal{C}$, there are $k$ elements that can be removed from $C$ to yield a set $D \in \Delta\mathcal{C}$. Thus, there are $|\mathcal{C}|k$ pairs $(C, D) \in \mathcal{C} \times \Delta\mathcal{C}$ such that $D \subseteq C$. If $D \in \Delta\mathcal{C}$, since $|D| = k-1$, it is possible to get $n-k+1$ sets $C$ such that $D \subseteq C$. Thus, $k|\mathcal{C}| \leq (n-k+1)|\Delta\mathcal{C}|$.

    The last two parts of this supplement show that the fraction of sets of size $k-1$ that are in the shadow of $\mathcal{C}$ and the fraction of sets of size $k+1$ that are in the shade of $\mathcal{C}$ are at least as large as the fraction of the size of sets $k$ in the collection $\mathcal{C}$. This fact is known as the *normalized matching property* of sets.

55. Let $M_n$ and $M_p$ be the multisets of prime divisors of the numbers $n$ and $p$, respectively, where $n, p \in \mathbb{N}$. Prove that $M_n + M_p = M_{np}$.

56. For two multisets $M$ and $P$ on a set $S$, denote by $M \leq P$ the fact that $M(x) \leq P(x)$ for every $x \in S$. Prove that $M \leq P$ implies $M \cup Q \leq P \cup Q$ and $M \cap Q \leq P \cap Q$ for every multiset $Q$ on $S$.

57. Let $M$ and $P$ be two multisets on a set $S$. Define the *multiset difference* $M - P$ by $(M - P)(x) = \max\{0, M(x) - P(x)\}$ for $x \in S$.
    a) Prove that $P \leq Q$ implies $M - P \geq M - Q$ and $P - M \leq Q - M$ for all multisets $M, P, Q$ on $S$.
    b) Prove that

$$M - (P \cup Q) = (M - P) \cap (M - Q),$$
$$M - (P \cap Q) = (M - P) \cup (M - Q),$$

    for all multisets $M, P, Q$ on $S$.
58. Define the symmetric difference of two multisets $M$ and $P$ as $(M \oplus P) = (M \cup P) - (M \cap P)$. Determine which properties of the symmetrical difference of sets can be extended to the symmetric difference of multisets.

## Bibliographical Comments

The reader may find [51] a useful reference for a detailed presentation of many aspects discussed in this chapter and especially for various variants of mathematical induction. Suggested introductory references to set theory are [62, 131].

# 2

# Algebras

## 2.1 Introduction

This chapter briefly presents several algebraic structures to the extent that they are necessary for the material presented in the subsequent chapters. Linear spaces and matrices, which are also discussed here, will receive an extensive treatment in the separate volume dedicated to linear algebra tools for data mining. We emphasize notions like operations, morphisms, and congruences that are of interest for the study of any algebraic structure.

## 2.2 Operations and Algebras

The notion of operation on a set is needed for introducing various algebraic structures on sets.

**Definition 2.1.** *Let $n \in \mathbb{N}$. An* $n$-ary operation on a set $S$ *is a function $f : S^n \longrightarrow S$. The number $n$ is the* arity *of the operation $f$.*

If $n = 0$, we have the special case of *zero-ary operations*. A zero-ary operation is a function $f : S^0 = \{\emptyset\} \longrightarrow S$, which is essentially a constant element of $S$, $f()$. Operations of arity 1 are referred to as *unary operations*.

*Binary operations* (of arity 2) are frequently used. For example, the union, intersection, and difference of subsets of a set $S$ are binary operations on the set $\mathcal{P}(S)$.

If $f$ is a binary operation on a set, we denote the result $f(x, y)$ of the application of $f$ to $x, y$ by $xfy$ rather than $f(x, y)$.

We now introduce certain important types of binary operations.

**Definition 2.2.** *An operation $f$ on a set $S$ is*
(i) associative *if $(xfy)fz = xf(yfz)$ for every $x, y, z \in S$,*
(ii) commutative *if $xfy = yfx$ for every $x, y, \in S$, and*
(iii) idempotent *if $xfx = x$ for every $x \in S$.*

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_2, © Springer-Verlag London Limited 2008

*Example 2.3.* Set union and intersection are both associative, commutative, and idempotent operations on every set of the form $\mathcal{P}(S)$.

The addition of real numbers "+" is an associative and commutative operation on $\mathbb{R}$; however, "+" is not idempotent.

The binary operation $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ given by $g(x, y) = \frac{x+y}{2}$ for $x, y \in \mathbb{R}$ is a commutative and idempotent operation of $\mathbb{R}$ that is not associative. Indeed, we have $(xgy)gz = \frac{x+y+2z}{4}$ and $xg(ygz) = \frac{2x+y+z}{4}$.

*Example 2.4.* The binary operations $\max\{x, y\}$ and $\min\{x, y\}$ are associative, commutative, and idempotent operations on the set $\mathbb{R}$.

Next, we introduce special elements relative to a binary operation on a set.

**Definition 2.5.** *Let $f$ be a binary operation on a set $S$.*
(i) *An element $u$ is a* unit for $f$ *if $xfu = ufx = x$ for every $x \in S$.*
(ii) *An element $z$ is a* zero for $f$ *if $zfu = ufz = z$ for every $x \in S$.*

Note that if an operation $f$ has a unit, then this unit is unique. Indeed, suppose that $u$ and $u'$ were two units of the operation $f$. According to Definition 2.5, we would have $ufx = xfu = x$ and, in particular, $ufu' = u'fu = u'$. Applying the same definition to $u'$ yields $u'fx = xfu' = x$ and, in particular, $u'fu = ufu' = u$. Thus, $u = u'$.

Similarly, if an operation $f$ has a zero, then this zero is unique. Suppose that $z$ and $z'$ were two zeros for $f$. Since $z$ is a zero, we have $zfx = xfz = z$ for every $x \in S$; in particular, for $x = z'$, we have $zfz' = z'fz = z$. Since $z'$ is zero, we also have $z'fx = xfz' = z'$ for every $x \in S$; in particular, for $x = z$, we have $z'fz = zfz' = z'$, and this implies $z = z'$.

**Definition 2.6.** *Let $f$ be a binary associative operation on $S$ such that $f$ has the unit $u$. An element $x$ has an* inverse *relative to $f$ if there exists $y \in S$ such that $xfy = yfx = u$.*

An element $x$ of $S$ has at most one inverse relative to $f$. Indeed, suppose that both $y$ and $y'$ are inverses of $x$. Then, we have

$$y = yfu = yf(xfy') = (yfx)fy' = ufy' = y',$$

which shows that $y$ coincides with $y'$.

If the operation $f$ is denoted by "+", then we will refer to the inverse of $x$ as the *additive inverse* of $x$, or the *opposite element* of $x$; similarly, when $f$ is denoted by "$\cdot$", we refer to the inverse of $x$ as the *multiplicative inverse of $x$*. The additive inverse of $x$ is usually denoted by $-x$, while the multiplicative inverse of $x$ is denoted by $x^{-1}$.

**Definition 2.7.** *Let $\mathcal{I} = \{f_i | i \in I\}$ be a set of operations on a set $S$ indexed by a set $I$. An* algebra type *is a mapping $\theta : I \longrightarrow \mathbb{N}$.*

*An* algebra of type $\theta$ *is a pair $\mathcal{A} = (A, \mathcal{I})$ such that*

(i) *$A$ is a set, and*
(ii) *the operation $f_i$ has arity $\theta(i)$ for every $i \in I$.*

*The algebra $\mathcal{A} = (A, \mathfrak{I})$ is* finite *if the set $A$ is finite. The set $A$ will be referred to as the* carrier *of the algebra $\mathcal{A}$.*

*If the indexing set $I$ is finite, we say that the type $\theta$ is a* finite type *and refer to $\mathcal{A}$ as an* algebra of finite type.

If $\theta : I \longrightarrow \mathbb{N}$ is a finite algebra type, we assume, in general, that the indexing set $I$ has the form $(0, 1, \ldots, n-1)$. In this case, we denote $\theta$ by the sequence $(\theta(0), \theta(1), \ldots, \theta(n-1))$.

Next, we discuss several algebra types.

**Definition 2.8.** *A* groupoid *is an algebra of type* $(2)$*, $\mathcal{A} = (A, \{f\})$. If $f$ is an associative operation, then we refer to this algebra as a* semigroup.

In other words, a groupoid is a set equipped with a binary operation $f$.

*Example 2.9.* The algebra $(\mathbb{R}, \{f\})$, where $f(x, y) = \frac{x+y}{2}$ is a groupoid. However, it is not a semigroup because $f$ is not an associative operation.

*Example 2.10.* Define the binary operation $g$ on $\mathbb{R}$ by $xgy = \ln(e^x + e^y)$ for $x, y \in \mathbb{R}$. Since

$$(xgy)gz = \ln(e^{xgy+e^z}) = \ln(e^x + e^y + e^z)$$
$$xg(ygz) = \ln(x + e^{ygz}) = \ln(e^x + e^y + e^z),$$

for every $x, y, z \in \mathbb{R}$ it follows that $g$ is an associative operation. Thus, $(\mathbb{R}, g)$ is a semigroup. It is easy to verify that this semigroup has no unit element.

**Definition 2.11.** *A* monoid *is an algebra of type* $(0, 2)$*, $\mathcal{A} = (A, \{e, f\})$, where $e$ is a zero-ary operation, $f$ is a binary operation, and $e$ is the unit element for $f$.*

*Example 2.12.* The algebras $(\mathbb{N}, \{1, \cdot\})$ and $(\mathbb{N}, \{0, \gcd\})$ are monoids. In the first case, the binary operation is the multiplication of natural numbers, the unit element is 1, and the algebra is clearly a monoid. In the second case, the binary operation $\gcd(m, n)$ yields the greatest common divisor of the numbers $m$ and $n$ and the unit element is 0.

We claim that gcd is an associative operation. Let $m, n, p \in \mathbb{N}$. We need to verify that $\gcd(m, \gcd(n, p)) = \gcd(\gcd(m, n), p)$.

Let $k = \gcd(m, \gcd(n, p))$. Then, $(k, m) \in \delta$ and $(k, \gcd(n, p)) \in \delta$, where $\delta$ is the divisibility relation introduced in Example 1.29. Since $\gcd(n, p)$ divides evenly both $n$ and $p$, it follows that $(k, n) \in \delta$ and $(k, p) \in \delta$. Thus, $k$ divides $\gcd(m, n)$, and therefore $k$ divides $h = \gcd(\gcd(m, n), p)$.

Conversely, $h$ being $\gcd(\gcd(m, n), p)$, it divides both $\gcd(m, n)$ and $p$. Since $h$ divides $\gcd(m, n)$, it follows that it divides both $m$ and $p$. Consequently, $h$ divides $\gcd(n, p)$ and therefore divides $k = \gcd(m, \gcd(n, p))$. Since

$k$ and $h$ are both natural numbers that divide each other evenly, it follows that $k = h$, which allows us to conclude that gcd is an associative operation. Since $n$ divides 0 evenly, for any $n \in \mathbb{N}$, it follows that $\gcd(0, n) = \gcd(n, 0) = n$, which shows that 0 is the unit for gcd.

**Definition 2.13.** *A* group *is an algebra of type* $(0, 2, 1)$*,* $\mathcal{A} = (A, \{e, f, h\})$*, where* $e$ *is a zero-ary operation,* $f$ *is a binary operation,* $e$ *is the unit element for* $f$*, and* $h$ *is a unary operation such that* $f(h(x), x) = f(x, h(x)) = e$ *for every* $x \in A$*.*

Note that if we have $xfy = yfx = e$, then $y = h(x)$. Indeed, we can write

$$h(x) = h(x)fe = h(x)f(xfy) = (h(x)fx)fy = efy = y.$$

We refer to the unique element $h(x)$ as the *inverse* of $x$. The usual notation for $h(x)$ is $x^{-1}$.

A special class of groups are the *Abelian groups*, also known as *commutative groups*. A group $\mathcal{A} = (A, \{e, f, h\})$ is Abelian if $xfy = yfx$ for all $x, y \in A$.

*Example 2.14.* The algebra $(\mathbb{Z}, \{0, +, -\})$ is an Abelian group, where "+" is the usual addition of integers, and the additive inverse of an integer $n$ is $-n$.

Traditionally, the binary operation of an Abelian group is denoted by "+".

**Definition 2.15.** *A* ring *is an algebra of type* $(0, 2, 1, 2)$*,* $\mathcal{A} = (A, \{e, f, h, g\})$*, such that* $\mathcal{A} = (A, \{e, f, h\})$ *is an Abelian group and* $g$ *is a binary associative operation such that*

$$xg(ufv) = (xgu)f(xgv),$$
$$(ufv)gx = (ugx)f(vgx),$$

*for every* $x, u, v \in A$*. These equalities are known as* left *and* right distributivity laws*, respectively.*

*The operation* $f$ *is known as the* ring addition*, while* $\cdot$ *is known as the* ring multiplication*. Frequently, these operations are denoted by "+" and "·", respectively.*

*Example 2.16.* The algebra $(\mathbb{Z}, \{0, +, -, \cdot\})$ is a ring. The distributivity laws amount to the well-known distributivity properties

$$p \cdot (q + r) = (p \cdot q) + (p \cdot r),$$
$$(q + r) \cdot p = (q \cdot p) + (r \cdot p),$$

for $p, q, r \in \mathbb{Z}$, of integer addition and multiplication.

*Example 2.17.* A more interesting type of ring can be defined on the set of numbers of the form $m+n\sqrt{2}$, where $m$ and $n$ are integers. The ring operations are given by

$$(m + n\sqrt{2}) + (p + q\sqrt{2}) = m + p + (n + q)\sqrt{2},$$
$$(m + n\sqrt{2}) \cdot (p + q\sqrt{2}) = m \cdot p + 2 \cdot n \cdot q + (m \cdot q + n \cdot p)\sqrt{2}.$$

If the multiplicative operation of a ring has a unit element 1, then we say that the ring is a *unitary ring*. We consider a unitary ring as an algebra of type $(0, 0, 2, 1, 2)$ by regarding the multiplicative unit as another zero-ary operation.

Observe, for example, that the ring $(\mathbb{Z}, \{0, 1, +, -, \cdot\})$ is a unitary ring. Also, note that the set of even numbers also generates a ring $(\{2k \mid k \in \mathbb{Z}\}, \{0, +, -, \cdot\})$. However, no multiplicative unit exists in this ring.

Rings with commutative multiplicative operations are known as *commutative rings*. All examples of rings considered so far are commutative rings. In Section 2.5, we shall see an important example of a noncommutative ring.

**Definition 2.18.** *A* field *is a pair* $\mathcal{A} = (A, \{e, f, h, g, u\})$ *such that* $\mathcal{A} = (A, \{e, f, h, g\})$ *is a commutative and unitary ring and* $u$ *is a unit for the binary operation* $g$ *such that every element* $x \neq e$ *has an inverse relative to the operation* $g$.

*Example 2.19.* The pair $\mathcal{R} = (\mathbb{R}, \{0, +, -, \cdot, 1\})$ is a field. Indeed, the multiplication "$\cdot$" is a commutative operation and 1 is a multiplicative unit. In addition, each element $x \neq 0$ has the inverse $\frac{1}{x}$.

## 2.3 Morphisms, Congruences, and Subalgebras

Morphisms are mappings between algebras of the same type that satisfy certain compatibility conditions with the operations of the type.

Let $\theta : I \longrightarrow \mathbb{N}$ be a type. To simplify notation, we denote the operations that correspond to the same element $i$ with the same symbol in every algebra of this type.

**Definition 2.20.** *Let* $\theta : I \longrightarrow \mathbb{N}$ *be a finite algebra type and let* $\mathcal{A} = (A, \mathfrak{I})$ *and* $\mathcal{B} = (B, \mathfrak{I})$ *be two algebras of the type* $\theta$. *A* morphism *is a function* $h : A \longrightarrow B$ *such that, for every operation* $f_i \in \mathfrak{I}$, *we have*

$$h(f_i(x_1, \ldots, x_{n_i})) = f_i(h(x_1), \ldots, h(x_{n_i}))$$

*for* $(x_1, \ldots, x_{n_i}) \in A^{n_i}$, *where* $n_i = \theta(i)$.

*If the algebras* $\mathcal{A}$ *and* $\mathcal{B}$ *are the same, then we refer to* $f$ *as an* endomorphism *of the algebra* $\mathcal{A}$.

*The set of morphisms between* $\mathcal{A}$ *and* $\mathcal{B}$ *is denoted by* $\mathsf{MOR}(\mathcal{A}, \mathcal{B})$.

*Example 2.21.* A morphism between the groupoids $\mathcal{A} = (A, \{f\})$ and $\mathcal{B} = (B, \{f\})$ is a mapping $h : A \longrightarrow B$ such that

$$h(f(x_1, x_2)) = f(h(x_1), h(x_2)) \tag{2.1}$$

for every $x_1$ and $x_2$ in $A$. Exactly the same definition is valid for semigroup morphisms.

If $\mathcal{A} = (A, \{e, f\})$ and $\mathcal{B} = (B, \{e, f\})$ are two monoids, where $e$ is a zero-ary operation and $f$ is a binary operation, then a morphism of monoids must satisfy the equalities $h(e) = e$ and $h(f(x_1, x_2)) = f(h(x_1), h(x_2))$ for $x_1, x_2 \in A$.

*Example 2.22.* Let $(\mathbb{N}, \{0, \gcd\})$ be the monoid introduced in Example 2.12. The function $h : \mathbb{N} \longrightarrow \mathbb{N}$ defined by $h(n) = n^2$ for $n \in \mathbb{N}$ is an endomorphism of this monoid because $\gcd(p, q)^2 = \gcd(p^2, q^2)$ for $p, q \in \mathbb{N}$.

*Example 2.23.* A morphism between two groups $\mathcal{A} = (A, \{e, \cdot, ^{-1}\})$ and $\mathcal{B} = (B, \{e, \cdot, ^{-1}\})$ satisfies the conditions $h(e) = e$, $h(x_1 \cdot x_2) = h(x_1) \cdot h(x_2)$, as well as $h(x_1^{-1}) = (h(x_1))^{-1}$, for $x_1, x_2 \in A$.

It is interesting to observe that, in the case of groups, the first and last conditions are consequences of the second condition, so they are superfluous. Indeed, choose $x_2 = e$ in the equality $h(x_1 \cdot x_2) = h(x_1) \cdot h(x_2)$; this yields $h(x_1) = h(x_1)h(e)$. By multiplying both sides with $h(x_1)^{-1}$ at the left and applying the associativity of the binary operation, we obtain $h(e) = e$. On the other hand, by choosing $x_2 = x_1^{-1}$, we have $e = h(e) = h(x_1)h(x_1^{-1})$, which implies $h(x_1^{-1}) = (h(x_1))^{-1}$.

*Example 2.24.* Let $\mathcal{A} = (A, \{0, +, ^{-1}, \cdot\})$ and $\mathcal{B} = (B, \{0, +, ^{-1}, \cdot\})$ be two rings. Then, $h : A \longrightarrow B$ is a ring morphism if $h(0) = 0$, $h(x_1 + x_2) = h(x_1) + h(x_2)$, and $h(x_1 \cdot x_2) = h(x_1) \cdot h(x_2)$.

**Definition 2.25.** *Let $\mathcal{A} = (A, \mathcal{I})$ be an algebra. An equivalence $\rho \in EQS(A)$ is a* congruence *if, for every operation $f$ of the algebra, $f : A^n \longrightarrow A$, $(x_i, y_i) \in \rho$ for $1 \leq i \leq n$ implies $(f(x_1, \ldots, x_n), f(y_1, \ldots, y_n)) \in \rho$ for $x_1, \ldots, x_n, y_1, \ldots, y_n \in A$.*

Recall that we introduce the kernel of a mapping in Definition 1.102. When the mapping $f$ is a morphism, we have further properties of $\mathbf{ker}(f)$.

**Theorem 2.26.** *Let $\mathcal{A} = (A, \mathcal{I})$ and $\mathcal{B} = (B, \mathcal{I})$ be two algebras of the type $\theta$ and let $h : A \longrightarrow B$ be a morphism. The relation $\boldsymbol{ker}(h)$ is a congruence of the algebra $\mathcal{A}$.*

*Proof.* Let $x_1, \ldots, x_n, y_1, \ldots, y_n \in A$ such that $(x_i, y_i) \in \mathbf{ker}(h)$ for $1 \leq i \leq n$; that is, $h(x_i) = h(y_i)$ for $1 \leq i \leq n$. By applying the definition of morphism, we can write for every $n$-ary operation in $\mathcal{I}$

$$\begin{aligned} h(f(x_1, \ldots, x_n)) &= f(h(x_1), \ldots, h(x_n)) \\ &= f(h(y_1), \ldots, h(y_n)) \\ &= h(f(y_1, \ldots, y_n)), \end{aligned}$$

which means that $(f(x_1, \ldots, x_n), f(y_1, \ldots, y_n)) \in \mathbf{ker}(f)$. Thus, $\mathbf{ker}(f)$ is a congruence. □

If $\mathcal{A} = (A, \mathcal{I})$ is an algebra and $\rho$ is a congruence of $\mathcal{A}$, then the quotient set $A/\rho$ (see Definition 1.115) can be naturally equipped with operations derived from the operations of $\mathcal{A}$ by

$$f([x_1]_\rho, \ldots, [x_n]_\rho) = [f(x_1, \ldots, x_n)]_\rho \tag{2.2}$$

for $x_1, \ldots, x_n \in A$. Observe first that the definition of the operation that acts on the $A/\rho$ is a correct one for if $y_i \in [x_i]_\rho$ for $1 \leq i \leq n$, then $[f(y_1, \ldots, y_n)]_\rho = [f(x_1, \ldots, x_n)]_\rho$.

**Definition 2.27.** *The* quotient algebra *of an algebra $\mathcal{A} = (A, \mathcal{I})$ and a congruence $\rho$ is the algebra $\mathcal{A}/\rho = (A/\rho, \mathcal{I})$, where each operation in $\mathcal{I}$ is defined starting from the corresponding operation $f$ in $\mathcal{A}$ by Equality (2.2).*

*Example 2.28.* Let $\mathcal{A} = (A, \{e, \cdot, ^{-1}\})$ be a group. An equivalence $\rho$ is a congruence if $(x_1, x_2), (y_1, y_2) \in \rho$ imply $(x_1^{-1}, x_2^{-1}) \in \rho$ and $(x_1 \cdot y_1, x_2 \cdot y_2) \in \rho$.

**Definition 2.29.** *Let $\mathcal{A} = (A, \mathcal{I})$ be an algebra. A subset $B$ of $A$ is* closed *if for every $n$-ary operation $f \in \mathcal{I}$, $x_1, \ldots, x_n \in B$ implies $f(x_1, \ldots, x_n) \in B$.*

Note that if the set $B$ is closed, then for every zero-ary operation $e$ of $\mathcal{I}$ we have $e \in B$.

Let $\mathcal{A} = (A, \mathcal{I})$ be an algebra and let $B$ be a closed subset of $A$. The pair $(B, \mathcal{I}')$, where $\mathcal{I}' = \{g_i = f_i \restriction_B \mid i \in I\}$, is an algebra of the same type as $\mathcal{A}$. We refer to it as a *subalgebra* of $\mathcal{A}$. Often we will refer to the set $B$ itself as a subalgebra of the algebra $\mathcal{A}$.

It is clear that the empty set is closed in an algebra $\mathcal{A} = (A, \mathcal{I})$ if and only if there is no zero-ary operation in $\mathcal{I}$.

We refer to subalgebras of particular algebras with more specific terms. For example, subalgebras of monoids or groups are referred to as *submonoids* or *subgroups*, respectively.

**Theorem 2.30.** *Let $\mathcal{A} = (A, \{e, \cdot, ^{-1}\})$ be a group. A nonempty subset $B$ of $A$ is a subgroup if and only if $x \cdot y^{-1} \in B$ for every $x, y \in B$.*

*Proof.* The necessity of the condition is immediate. To prove that the condition is sufficient, observe that since $B \neq \emptyset$ there is $x \in B$, so $x \cdot x^{-1} = e \in B$.

Next, let $x \in B$. Since $e \cdot x^{-1} = x^{-1}$ it follows that $x^{-1} \in B$. Finally, if $x, y \in B$, then $x \cdot y = x \cdot (y^{-1})^{-1} \in B$, which shows that $B$ is indeed a subgroup. □

*Example 2.31.* Let $\mathcal{A} = (A, \{e, \cdot, ^{-1}\})$ be a group. For $u \in A$, define the set $C_u = \{x \in G \mid xu = ux\}$. If $x \in C_u$, then $xu = ux$, which implies $xux^{-1} = u$, so $ux^{-1} = x^{-1}u$. Thus, $x^{-1} \in C_u$. It is easy to see that $e \in C_u$ and $x, y \in C_u$ implies $xy \in C_u$. Thus $C_u$ is a subgroup.

## 2.4 Linear Spaces

Linear spaces are studied in a distinct mathematical discipline, named *linear algebra*. Applications of linear spaces in data mining will be the object of a dedicated volume. In this section, we discuss basic properties of linear spaces that are useful for the present volume.

**Definition 2.32.** *Let $L$ be a nonempty set and let $\mathcal{F} = (F, \{0, +, -, \cdot, \})$ be a field whose carrier is a set $F$. An* $\mathcal{F}$-linear space *is a triple $(L, +, \cdot)$ such that $(L, \{\boldsymbol{0}, +, -\})$ is an Abelian group and $\cdot : F \times L \longrightarrow L$ is an operation such that the following conditions are satisfied*

(i) $a \cdot (b \cdot \boldsymbol{x}) = (a \cdot b) \cdot \boldsymbol{x}$,
(ii) $1 \cdot \boldsymbol{x} = \boldsymbol{x}$,
(iii) $a \cdot (\boldsymbol{x} + \boldsymbol{y}) = a \cdot \boldsymbol{x} + a \cdot \boldsymbol{y}$, *and*
(iv) $(a + b) \cdot \boldsymbol{x} = a \cdot \boldsymbol{x} + b \cdot \boldsymbol{x}$

*for every $a, b \in F$ and $\boldsymbol{x}, \boldsymbol{y} \in L$.*

If $\mathcal{F}$ is the field of real numbers $\mathcal{R}$, then we will refer to any $\mathcal{R}$-linear space as a *real linear space*.

Note that the commutative binary operation of $L$ is denoted by the same symbol "+" as the corresponding operation of the field $F$. The operation $\cdot : F \times L \longrightarrow L$ is an external operation since its two arguments belong to two different sets, $F$ and $L$. Again, this operation is denoted by the same symbol used for denoting the multiplication on $F$.

The elements of the set $L$ will be denoted using bold letters $\mathbf{x}, \mathbf{y}, \mathbf{z}$, etc. The members of the field will be denoted by small letters from the beginning of the alphabet.

The additive element $\mathbf{0}$ is a special element called the *zero element*; every $\mathcal{F}$-linear space must contain at least this element.

*Example 2.33.* The set $\mathbb{R}^n$ of $n$-tuples of real numbers is an $\mathcal{R}$-linear space under the definitions

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1 + y_1, \ldots, x_n + y_n), \\ a \cdot \mathbf{x} &= (a \cdot x_1, \ldots, a \cdot x_n), \end{aligned}$$

of the operations $+$ and $\cdot$, where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$. In this linear space, the zero of the Abelian group is the $n$-tuple $\mathbf{0} = (0, \ldots, 0)$.

*Example 2.34.* Let $S$ be a set. The set of real-valued functions defined on $S$ is a real linear space. The addition of functions is given by $(f + g)(s) = f(s) + g(s)$, and the multiplication of a function with a real number is defined by $(af)(s) = af(s)$ for $s \in S$ and $a \in \mathbb{R}$.

*Example 2.35.* Let $C$ be the set of real-valued continuous functions defined on $\mathbb{R}$,

$$C = \{f : \mathbb{R} \longrightarrow \mathbb{R} \mid f \text{ is continuous}\}.$$

Define $f + g$ by $(f + g)(x) = f(x) + g(x)$ and $(a \cdot f)(x) = a \cdot f(x)$ for $x \in \mathbb{R}$.

The triple $(C, +, \cdot)$ is a real linear space.

**Definition 2.36.** *Let $\mathcal{F} = (F, \{0, +, -, \cdot, \})$ be a field and let $\mathfrak{L} = (L, +, \cdot)$ be an $\mathcal{F}$-linear space. For a finite subset $K = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of $L$, a* linear combination of $K$ *is a member of $L$ of the form $c_1\boldsymbol{x}_1 + \cdots + c_n\boldsymbol{x}_n$, where $c_1, \ldots, c_n \in F$.*

*A subset $K = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of $L$ is* linearly independent *if $c_1\boldsymbol{x}_1 + \cdots + c_n\boldsymbol{x}_n = \boldsymbol{0}$ implies $c_1 = \cdots = c_n = 0$. If $K$ is not linearly independent, we refer to $K$ as a* linearly dependent *set.*

If $\mathbf{x} \neq \mathbf{0}$, then the set $\{\mathbf{x}\}$ is linearly independent. Of course, the set $\{\mathbf{0}\}$ is not linearly independent because $1\mathbf{0} = \mathbf{0}$.

It is easy to see that if $K$ is a linearly independent subset of a linear space, then any subset of $K$ is linearly independent.

*Example 2.37.* Let $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots)$ be a binary vector that has a unique nonzero component in place $i$, where $1 \leq i \leq n$. The set $E = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is linearly independent. Indeed, suppose that $c_1\mathbf{e}_1 + \cdots + c_n\mathbf{e}_n = \mathbf{0}$. This is equivalent to $(c_1, \ldots, c_n) = (0, \ldots, 0)$, that is, $c_1 = \cdots = c_n = 0$. Thus, $E$ is linearly independent.

If $U$ is an arbitrary subset of a linear space, we say that $\mathbf{x} \in L$ is a *linear combination of $U$* if there exists a finite subset $K$ of $U$ such that $\mathbf{x}$ is a linear combination of $K$.

**Theorem 2.38.** *Let $(L, +, \cdot)$ be an $\mathcal{F}$-linear space. A subset $K$ of $L$ is linearly independent if and only if for every $\boldsymbol{x} \in L$ that is a linear combination of $K$, $\boldsymbol{x} = \sum_i c_i\boldsymbol{x}_i$, the coefficients $c_i$ are uniquely determined.*

*Proof.* Suppose that $\mathbf{x} = c_1\mathbf{x}_1 + \cdots + c_n\mathbf{x}_n = c'_1\mathbf{x}_1 + \cdots + c'_n\mathbf{x}_n$ and there exists $i$ such that $c_i \neq c'_i$. This implies $\sum_{i=1}^{n}(c_i - c'_i)\mathbf{x}_i = \mathbf{0}$, which contradicts the linear independence of $K$. □

**Definition 2.39.** *A subset $S$ of a linear space $(L, +, \cdot)$* spans *the space $L$ (or $S$ generates the linear space) if every $\boldsymbol{x} \in L$ can be written as a linear combination of $S$.*

*A* basis *of the linear space $(L, +, \cdot)$ is a linearly independent subset that spans the linear space.*

In view of Theorem 2.38, a set $B$ is a basis if every $\mathbf{x} \in L$ can be written uniquely as a linear combination of elements of $B$.

**Definition 2.40.** *Let $\mathcal{F} = (F, \{0, +, -, \cdot, \})$ be a field. A* subspace *of a $\mathcal{F}$-linear space $(L, +, \cdot)$ is a nonempty subset $U$ of $L$ such that $\boldsymbol{x}, \boldsymbol{y} \in U$ implies $\boldsymbol{x} + \boldsymbol{y} \in U$ and $a \cdot \boldsymbol{x} \in U$ for every $a \in F$.*

Note that the set $U_0 = \{\mathbf{0}\}$ is a subspace of any $\mathcal{F}$-linear space $(L, +, \cdot)$. Moreover, $U_0$ is included in any subspace of the linear space.

If $\{K_i \mid i \in I\}$ is a nonempty collection of subspaces of a linear space, then $\bigcap\{K_i \mid i \in I\}$ is also a linear subspace.

**Theorem 2.41.** *Let $\mathfrak{L} = (L, +, \cdot)$ be an $\mathcal{F}$-linear space. The following statements are equivalent:*

(i) *The finite set $K = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is spanning the linear space $(L, +, \cdot)$ and $K$ is minimal with this property.*

(ii) *$K$ is a finite basis for $(L, +, \cdot)$.*

(iii) *The finite set $K$ is linearly independent, and $K$ is maximal with this property.*

*Proof.* (i) implies (ii): We need to prove that $K$ is linearly independent. Suppose that this is not the case. Then, there exist $c_1, \ldots, c_n \in F$ such that $c_1\mathbf{x}_1 + \cdots + c_n\mathbf{x}_n = \mathbf{0}$ and at least one of $c_1, \ldots, c_n$, say $c_i$, is nonzero. Then, $\mathbf{x}_i = -\frac{c_1}{c_i}\mathbf{x}_1 - \cdots - \frac{c_n}{c_i}\mathbf{x}_n$, and this implies that $K - \{\mathbf{x}_i\}$ also spans the linear space, thus contradicting the minimality of $K$.

(ii) implies (i): Let $K$ be a finite basis. Suppose that $K'$ is a proper subset of $K$ that spans $L$. Then, if $\mathbf{z} \in K - K'$, $z'$ is a linear combination of elements of $K'$, which contradicts the fact that $K$ is a basis.

We leave to the reader the proof of the equivalence between (ii) and (iii). □

**Corollary 2.42.** *Every linear space that is spanned by a finite subset has a finite basis. Further, if $B$ is a finite basis for an $\mathcal{F}$-linear space $(L, +, \cdot)$, then each finite subset $U$ of $L$ such that $|U| = |B| + 1$ is linearly dependent.*

*Proof.* This statement follows directly from Theorem 2.41. □

**Corollary 2.43.** *If $B$ and $B'$ are two finite bases for a linear space $(L, +, \cdot)$, then $|B| = |B'|$.*

*Proof.* If $B$ is a finite basis, then $|B|$ is the maximum number of linearly independent elements in $L$. Thus, $|B'| \leq |B|$. Reversing the roles of $B$ and $B'$, we obtain $|B| \leq |B'|$, so $|B| = |B'|$. □

Thus, the number of elements of a finite basis of $L$ is a characteristic of $L$ and does not depend on any particular basis.

**Definition 2.44.** *A linear space $(L, +, \cdot)$ is $n$-dimensional if there exists a basis of $L$ such that $|B| = n$. The number $n$ is the dimension of $L$ and is denoted by* $\dim(L)$.

**Definition 2.45.** *An* inner product *on a real linear space $(L, +, \cdot)$ is a function $p : L^2 \longrightarrow \mathbb{R}$ that has the following properties*

(i) $p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y}, \boldsymbol{x})$,

(ii) $p(\boldsymbol{x}, a\boldsymbol{y}) = ap(\boldsymbol{x}, \boldsymbol{y})$,
(iii) $p(\boldsymbol{x}, \boldsymbol{y} + \boldsymbol{y}') = p(\boldsymbol{x}, \boldsymbol{y}) + p(\boldsymbol{x}, \boldsymbol{y}')$, *and*
(iv) $p(\boldsymbol{x}, \boldsymbol{x}) \geq 0$ *and* $p(\boldsymbol{x}, \boldsymbol{x}) = 0$ *implies* $\boldsymbol{x} = \boldsymbol{0}$,
*for every* $\boldsymbol{x}, \boldsymbol{y} \in L$ *and* $a \in \mathbb{R}$.

The linear product $p(\mathbf{x}, \mathbf{y})$ is denoted by $\mathbf{x} \cdot \mathbf{y}$.

*Example 2.46.* An inner product on $\mathbb{R}^n$ is defined by

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \cdots + x_n y_n$$

for every $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$.

**Definition 2.47.** *A* norm *on a real linear space* $(L, +, \cdot)$ *is a mapping* $\nu : L \longrightarrow \mathbb{R}_{\geq 0}$ *such that*
(i) $\nu(\boldsymbol{x}) \geq 0$, *and* $\nu(\boldsymbol{x}) = 0$ *implies* $\boldsymbol{x} = \boldsymbol{0}$,
(ii) $\nu(a\boldsymbol{x}) = |a|\nu(\boldsymbol{x})$, *and*
(iii) $\nu(\boldsymbol{x} + \boldsymbol{y}) \leq \nu(\boldsymbol{x}) + \nu(\boldsymbol{y})$
*for every* $a \in \mathbb{R}$ *and* $\boldsymbol{x}, \boldsymbol{y} \in L$.

**Theorem 2.48.** *Let* $(L, +, \cdot)$ *be a linear space and let* $\boldsymbol{x}, \boldsymbol{y} \in L$. *We have the inequality*

$$\Big|\nu(\boldsymbol{x}) - \nu(\boldsymbol{y})\Big| \leq \nu(\boldsymbol{x} - \boldsymbol{y}).$$

*Proof.* By applying the definition of the norms, we can write

$$\nu(\mathbf{x}) \leq \nu(\mathbf{x} - \mathbf{y}) + \nu(\mathbf{y}),$$
$$\nu(\mathbf{y}) \leq \nu(\mathbf{y} - \mathbf{x}) + \nu(\mathbf{x}),$$

which are equivalent to

$$\nu(\mathbf{x}) - \nu(\mathbf{y}) \leq \nu(\mathbf{x} - \mathbf{y}) \text{ and}$$
$$\nu(\mathbf{y}) - \nu(\mathbf{x}) \leq \nu(\mathbf{y} - \mathbf{x}),$$

respectively.

Since $\nu(\mathbf{x} - \mathbf{y}) = \nu(\mathbf{y} - \mathbf{x})$ (by the second property of Definition 2.47), it follows that

$$-\nu(\mathbf{x} - \mathbf{y}) \leq \nu(\mathbf{x}) - \nu(\mathbf{y}) \leq \nu(\mathbf{x} - \mathbf{y}),$$

which gives the desired inequality. □

**Theorem 2.49 (Cauchy's Inequality).** *Let* $(L, +, \cdot)$ *be a linear space. For every* $\boldsymbol{x}, \boldsymbol{y} \in L$, *we have*

$$(\boldsymbol{x} \cdot \boldsymbol{y})^2 \leq (\boldsymbol{x} \cdot \boldsymbol{x})(\boldsymbol{y} \cdot \boldsymbol{y}),$$

*where* $\boldsymbol{u} \cdot \boldsymbol{v}$ *denotes the inner product of* $\boldsymbol{u}, \boldsymbol{v} \in L$.

*Proof.* Consider the function $f : \mathbb{R} \longrightarrow \mathbb{R}$ defined by $f(\lambda) = (\mathbf{x}+\lambda\mathbf{y})\cdot(\mathbf{x}+\lambda\mathbf{y})$ for $\lambda \in \mathbb{R}$. Observe that $f(\lambda) = \mathbf{x}\cdot\mathbf{x} + 2\lambda\mathbf{x}\cdot\mathbf{y} + \lambda^2\mathbf{y}\cdot\mathbf{y}$. By Definition 2.45, we have $f(\lambda) \geq 0$ for every $\lambda$ and this implies that the discriminant of the quadratic expression, $(\mathbf{x}\cdot\mathbf{y})^2 - (\mathbf{x}\cdot\mathbf{x})(\mathbf{y}\cdot\mathbf{y})$ must be not greater than 0. This gives the desired inequality. □

If an inner product exists on a real linear space $(L, +, \cdot)$, a norm can be defined by

$$\nu(x) = \sqrt{\mathbf{x}\cdot\mathbf{x}},$$

for every $\mathbf{x} \in L$. We leave it to the reader to verify that $\nu$ is indeed a norm. The satisfaction of the third condition of Definition 2.47 follows immediately from Cauchy's inequality.

An alternative notation for the norm of an element $\mathbf{x} \in L$ is $\| \mathbf{x} \|$.

*Example 2.50.* The norm $\| \mathbf{x} \|_2$ on $\mathbb{R}^n$ defined by

$$\| \mathbf{x} \|_2 = \sqrt{x_1^2 + \cdots + x_n^2},$$

where $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, is induced by the inner product on $\mathbb{R}^n$ defined in Example 2.46 because $\| \mathbf{x} \|_2 = \sqrt{\mathbf{x}\cdot\mathbf{x}}$.

This norm is known as the *Euclidean norm on* $\mathbb{R}^n$.

There exist norms in linear spaces that cannot be generated by inner products (see Exercises 18 and 19).

**Definition 2.51.** *Let* $\boldsymbol{w}$ *be a vector in* $\mathbb{R}^n$ *and let* $t \in \mathbb{R}$. *A* hyperplane *in* $\mathbb{R}^n$ *is a subset* $H_{\boldsymbol{w},t}$ *of* $\mathbb{R}^n$ *defined by*

$$H_{\boldsymbol{w},t} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{w}\boldsymbol{x} = t\}.$$

*The vector* $\boldsymbol{w}$ *is said to be* normal *to the hyperplane* $H_{\boldsymbol{w},t}$.

## 2.5 Matrices

We define a class of two-argument functions that is ubiquitous in mathematics and is very important for the applications that we consider here.

**Definition 2.52.** *Let* $S$ *be a nonempty set. A* matrix on $S$ *is a function*

$$M : \{1, \ldots, m\} \times \{1, \ldots, n\} \longrightarrow S.$$

*The pair* $(m, n)$ *is the* format *of the matrix* $M$.

Matrices can be conceived as two-dimensional arrays as shown below:

$$\begin{pmatrix} M(1,1) & M(1,2) & \dots & M(1,n) \\ M(2,1) & M(2,2) & \dots & M(2,n) \\ \vdots & \vdots & \dots & \vdots \\ M(i,1) & M(i,2) & \dots & M(i,n) \end{pmatrix}.$$

Alternatively, a matrix $M : \{1,\dots,m\} \times \{1,\dots,n\} \longrightarrow S$ can be regarded as consisting of $m$ rows, where each row is a sequence of the form

$$(M(i,1), M(i,2), \dots, M(i,n)),$$

for $1 \leq i \leq n$, or as a collection of $n$ columns of the form

$$\begin{pmatrix} M(1,j) \\ M(2,j) \\ \vdots \\ M(m,j) \end{pmatrix},$$

where $1 \leq j \leq m$.

If $M : \{1,\dots,m\} \times \{1,\dots,n\} \longrightarrow S$ is a matrix on $S$, we shall say that $M$ is an $(m \times n)$-matrix on $S$. The set of all such matrices will be denoted by $S^{m \times n}$.

*Example 2.53.* Let $S = \{0, 1\}$. The matrix

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

is a $(3 \times 2)$-matrix on the set $S$.

The element $M(i,j)$ of the matrix $M$ will usually be denoted by $M_{ij}$.

Regardless of the set $S$, we can consider the following two types of matrices over $S$.

**Definition 2.54.** *A* square matrix *on $S$ is an $(n \times n)$-matrix on the set $S$ for some $n \geq 1$.*

*An $(n \times n)$-square matrix on $S$ is* symmetric *if $M_{ij} = M_{ji}$ for every $i, j$ such that $1 \leq i, j \leq n$.*

*Example 2.55.* The $(3 \times 3)$-matrix

$$\begin{pmatrix} 1 & 0.5 & 1 \\ 0.5 & 1 & 2 \\ 1 & 2 & 0.3 \end{pmatrix}$$

over the set of reals $\mathbb{R}$ is symmetric.

**Definition 2.56.** *The* transpose of an $(m \times n)$-matrix $M$ *is an $(n \times m)$-matrix $M^{tran}$ defined by $M^{tran}_{ij} = M_{ji}$.*

In other words, the transpose $M^{tran}$ of a matrix $M$ has as rows the transposed columns of $M$; equivalently, the columns of $M^{tran}$ are the transposed rows of $M$.

It is easy to verify that, for any matrix $M \in S^{m\times n}$, we have

$$(M^{tran})^{tran} = M.$$

If the set $S$ is equipped with the structure of a unitary ring $(S, \{0, +, -, \cdot\})$, then we can consider more interesting facts on the set $S^{m\times n}$.

**Definition 2.57.** *The* $(n \times n)$-unit matrix *on the ring* $(S, \{0, +, -, \cdot\})$ *is the square matrix* $I_n \in S^{n\times n}$ *given by*

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

*whose entries located outside its main diagonal are* 0*s.*

*The* $(m \times n)$-zero matrix *is the* $(m \times n)$*-matrix* $O_{m,n} \in S^{n\times n}$ *given by*

$$O_{m,n} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

The ring structure $(S, \{0, +, -, \cdot\})$ allows the definition of matrix addition and matrix multiplication.

**Definition 2.58.** *Let* $(S, \{0, +, -, \cdot\})$ *be a ring and let* $M, P \in S^{m\times n}$ *be two matrices that have the same format. The* sum *of the matrices* $M$ *and* $P$ *is the matrix* $M + P$ *having the same format and defined by*

$$(M + P)_{ij} = M_{ij} + P_{ij}$$

*for* $1 \le i \le m$ *and* $1 \le j \le n$.

*Example 2.59.* Let $M, P \in \mathbb{R}^{2\times 3}$ be two matrices given by

$$M = \begin{pmatrix} 1 & -2 & 3 \\ 0 & 2 & -1 \end{pmatrix} \text{ and } P = \begin{pmatrix} -1 & 2 & 3 \\ 1 & 4 & 2 \end{pmatrix}.$$

Their sum is the matrix

$$M + P = \begin{pmatrix} 0 & 0 & 6 \\ 1 & 6 & 1 \end{pmatrix}.$$

It is easy to verify that the matrix sum is an associative and commutative operation on $S^{m\times n}$; that is,

$$M + (P + Q) = (M + P) + Q,$$
$$M + P = P + M,$$

for all $M, P, Q \in S^{m\times n}$.

The zero matrix $O_{m,n}$ acts as an additive unit on the set $S^{m\times n}$; that is,

$$M + O_{m,n} = O_{m,n} + M$$

for every $M \in S^{m\times n}$.

The additive inverse, or the opposite of a matrix $M \in S^{m\times n}$, is the matrix $-M$ given by $(-M)_{ij} = -M_{ij}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$.

*Example 2.60.* The opposite of $M \in \mathbb{R}^{2\times 3}$, given by

$$M = \begin{pmatrix} 1 & -2 & 3 \\ 0 & 2 & -1 \end{pmatrix}$$

is the matrix

$$-M = \begin{pmatrix} -1 & 2 & -3 \\ 0 & -2 & 1 \end{pmatrix}.$$

It is immediate that $M + (-M) = O_{2,3}$.

The discussion above shows that the set of matrices $S^{m\times n}$ defined on a ring $(S, \{0, +, -, \cdot\})$ is an Abelian group $(S^{m\times n}, \{O_{m,n}, +, -\})$.

**Definition 2.61.** *Let $(S, \{0, +, -, \cdot\})$ be a ring and let $M \in S^{m\times n}$ and $P \in S^{n\times p}$ be two matrices. The* product *of the matrices $M, P$ is the matrix $Q \in S^{m\times p}$ defined by*

$$Q_{ik} = \sum_{j=1}^{n} M_{ij} P_{jk},$$

*where $1 \leq i \leq m$ and $1 \leq k \leq p$. The product of the matrices $M, P$ will be denoted by $MP$.*

The matrix product is a partial operation because in order to multiply two matrices $M$ and $P$, they must have the formats $m \times n$ and $n \times p$, respectively. In other words, the number of columns of the first matrix must equal the number of rows of the second matrix.

**Theorem 2.62.** *Matrix multiplication is associative.*

*Proof.* Let $M \in S^{m\times n}$, $P \in S^{n\times p}$, and $R \in S^{p\times r}$ be three matrices, where $(S, \{0, +, -, \cdot\})$ is a ring. We need to prove that $(MP)R = M(PR)$.

By applying the definition of the matrix product, we have

$$\begin{aligned}((MP)R)_{i\ell} &= \sum_{k=1}^{p}(MP)_{ik}R_{k\ell}\\ &= \sum_{k=1}^{p}\left(\sum_{j=1}^{n}M_{ij}P_{jk}\right)R_{k\ell}\\ &= \sum_{j=1}^{n}M_{ij}\sum_{k=1}^{p}P_{jk}R_{k\ell}\\ &= \sum_{j=1}^{n}M_{ij}(PR)_{j\ell}\\ &= (M(PR))_{i\ell}\end{aligned}$$

for $1 \leq i \leq m$ and $1 \leq \ell \leq r$, which shows that matrix multiplication is indeed associative. □

**Theorem 2.63.** *If $M \in S^{m\times n}$, then $I_m M = MI_n = M$.*

*Proof.* The statement follows immediately from the definition of a matrix product. □

Note that if $M \in S^{n\times n}$, then $I_n M = MI_n = M$, so $I_n$ is a unit relative to matrix multiplication considered as an operation on the set of square matrices $S^{n\times n}$.

The product of matrices is not commutative. Indeed, consider the matrices $M, P \in \mathbb{Z}^{2\times 2}$ defined by

$$M = \begin{pmatrix}0 & 1\\ 2 & 3\end{pmatrix} \text{ and } P = \begin{pmatrix}-1 & 1\\ 1 & 0\end{pmatrix}.$$

We have

$$MP = \begin{pmatrix}1 & 0\\ 1 & 2\end{pmatrix} \text{ and } P = \begin{pmatrix}2 & 2\\ 0 & 1\end{pmatrix},$$

so $MP \neq PM$.

Matrices are used in studying partitions.

**Definition 2.64.** *Let $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$ of a finite set $S = \{s_1, \ldots, s_\ell\}$. The* contingency matrix *of the partitions $\pi$ and $\sigma$ is the $(m \times n)$-matrix $Q(\pi, \sigma)$, where $Q(\pi, \sigma)_{ij} = |B_i \cap C_j|$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. The element $Q(\pi, \sigma)_{ij}$ will be denoted by $q_{ij}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$.*

Starting from a contingency matrix $Q(\pi, \sigma)$, we introduce the *marginal totals* of such a matrix:

$$q_{\cdot j} = \sum_{i=1}^{m} q_{ij} \text{ for } 1 \leq j \leq n \text{ and}$$
$$q_{i\cdot} = \sum_{j=1}^{n} q_{ij} \text{ for } 1 \leq i \leq m.$$

Clearly, $|C_j| = q_{\cdot j}$, $|B_i| = q_{i\cdot}$, and $|S| = \sum_{i=1}^{m} q_{i\cdot} = \sum_{j=1}^{n} q_{\cdot j} = |S|$. Also, we have

$$\sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij} = \sum_{i=1}^{m} q_{i\cdot} = \sum_{j=1}^{n} q_{\cdot j} = \ell.$$

Let $\rho_\pi$ and $\rho_\sigma$ be the equivalence relations that correspond to the partitions $\pi$ and $\sigma$, introduced in Theorem 1.116.

The set of unordered pairs of elements of $S$ was denoted by $\mathcal{P}_2(S)$. If $|S| = \ell$, then it is easy to see that $|\mathcal{P}_2(S)| = \frac{\ell^2-\ell}{2}$ distinct unordered pairs of elements. An unordered pair $\{s, s'\}$ belongs to one of the following four classes:

1. *Type 1* pairs are those pairs such that $(s, s') \in \rho_\pi$ and $(s, s') \in \rho_\sigma$.
2. *Type 2* pairs are those pairs such that $(s, s') \notin \rho_\pi$ and $(s, s') \notin \rho_\sigma$.
3. *Type 3* pairs are those pairs such that $(s, s') \notin \rho_\pi$ and $(s, s') \in \rho_\sigma$.
4. *Type 4* pairs are those pairs such that $(s, s') \in \rho_\pi$ and $(s, s') \notin \rho_\sigma$.

The *number of agreements* $agr(\pi, \sigma)$ of the partitions $\pi$ and $\sigma$ is the total number of pairs of types 1 and 2; the number of disagreements of these partitions $dagr(\pi, \sigma)$ is the total number of pairs of types 3 and 4. Clearly, we have

$$agr(\pi, \sigma) + dagr(\pi, \sigma) = \frac{\ell^2 - \ell}{2}.$$

Note that the number of pairs of type 1 equals

$$\ell_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{q_{ij}^2 - q_{ij}}{2} = \frac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}^2 - \ell\right).$$

The number of pairs of type 3 is

$$\ell_3 = \sum_{j=1}^{n} \frac{q_{\cdot j}^2 - q_{\cdot j}}{2} - \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{q_{ij}^2 - q_{ij}}{2} = \frac{1}{2}\left(\sum_{j=1}^{n} q_{\cdot j}^2 - \sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}^2\right),$$

while the number of pairs of type 4 is:

$$\ell_4 = \sum_{i=1}^{m} \frac{q_{i\cdot}^2 - q_{i\cdot}}{2} - \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{q_{ij}^2 - q_{ij}}{2} = \frac{1}{2}\left(\sum_{i=1}^{m} q_{i\cdot}^2 - \sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}^2\right).$$

The equalities above allow us to compute the number of pairs of type 2 as

$$\begin{aligned}\ell_2 &= \frac{\ell^2-\ell}{2} - \ell_1 - \ell_3 - \ell_4 \\ &= \frac{1}{2}\left(\ell^2 + \sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}^2 - \sum_{i=1}^{m} q_{i\cdot}^2 - \sum_{j=1}^{n} q_{\cdot j}^2\right).\end{aligned}$$

Thus, we obtain

$$agr(\pi,\sigma) = \frac{1}{2}\left(2\sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}^2 + \ell^2 - \ell - \sum_{i=1}^{m} q_{i\cdot}^2 - \sum_{j=1}^{n} q_{\cdot j}^2\right), \tag{2.3}$$

$$dagr(\pi,\sigma) = \frac{1}{2}\left(\sum_{j=1}^{n} q_{\cdot j}^2 + \sum_{i=1}^{m} q_{i\cdot}^2 - 2\sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}^2\right). \tag{2.4}$$

## Exercises and Supplements

1. Let $a, b, c, d$ be four real numbers and let $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be the binary operation on $\mathbb{R}$ defined by

   $$f(x,y) = axy + bx + cy + d$$

   for $x, y \in \mathbb{R}$.
   a) Prove that $f$ is a commutative operation if and only if $b = c$.
   b) Prove that $f$ is an idempotent operation if and only if $a = d = 0$ and $b + c = 1$.
   c) Prove that $f$ is an associative operation if and only if $b = c$ and $b^2 - b - ad = 0$.
2. Let $*$ be an operation defined on a set $T$ and let $f : S \longrightarrow T$ be a bijection. Define the operation $\circ$ on $S$ by $x \circ y = f^{-1}(f(x) * f(y))$ for $x, y \in T$. Prove that:
   a) The operation $\circ$ is commutative (associative) if and only if $*$ is commutative (associative).
   b) If $u$ is a unit element for $*$, then $v = f^{-1}(u)$ is a unit element for $\circ$.
3. Prove that the algebra $(\mathbb{R}_{>0}, \{*\})$, where $*$ is a binary operation defined by $x * y = x^{\log y}$, is a commutative semigroup.
4. Define the binary operation $\circ$ on $\mathbb{Z} \times \mathbb{Z}$ by $(x_1, y_1) \circ (x_2, y_2) = (x_1x_2 + 2y_1y_2, x_1y_2 + x_2y_1)$. Prove that the algebra $(\mathbb{Z} \times \mathbb{Z}, \{\circ\})$ is a commutative monoid.
5. Let $\mathcal{A} = (A, \{e, \cdot, ^{-1}\})$ be a group and let $\rho$ be a congruence of $\mathcal{A}$. Prove that the set $\{x \in A \mid (x, e) \in \rho\}$ is a subalgebra of $\mathcal{A}$, that is, a subgroup.

Let $S$ be a set and let $(G, \{u, \cdot, ^{-1}\})$ be a group. A *group action on* $S$ is a binary function $f : G \times S \longrightarrow S$ that satisfies the following conditions:
(i) $f(x \cdot y, s) = f(x, f(y, s))$ for all $x, y \in G$ and $s \in S$;
(ii) $f(u, s) = s$ for every $s \in S$.
The element $f(x, s)$ is denoted by $xs$. If an action of a group on a set $S$ is defined, we say that the group is acting on the set $S$.

The *orbit* of an element $s \in S$ is the subset $O_s$ of $S$ defined by $O_s = \{xs \mid x \in G\}$. The *stabilizer* of $s$ is the subset of $G$ given by $T_x = \{x \in G \mid xs = s\}$.

6. Let $(G, \{u, \cdot, ^{-1}\})$ be a group acting on a set $S$. Prove that if $O_s \cap O_z \neq \emptyset$, then $O_s = O_z$.
7. Let $(G, \{u, \cdot, ^{-1}\})$ be a group acting on a set $S$. Prove that:
   a) If $O_s \cap O_z \neq \emptyset$, then $O_s = O_z$ for every $s, z \in S$.
   b) For every $s \in S$, the stabilizer of $s$ is a subgroup of $G$.
8. Let $B$ be a subgroup of a group $\mathcal{A} = (A, \{e, \cdot, ^{-1}\})$. Prove that:
   a) The relations $\rho_B$ and $\sigma_B$ defined by

$$\rho_B = \{(x, y) \in A \times A \mid x \cdot y^{-1} \in B\},$$
$$\sigma_B = \{(x, y) \in A \times A \mid x^{-1} \cdot y \in B\},$$

   are equivalence relations on $A$.
   b) $(x, y) \in \rho_B$ implies $(x \cdot z, y \cdot z) \in \rho_B$ and $(x, y) \in \sigma_B$ implies $(z \cdot x, z \cdot y) \in \sigma_B$ for every $z \in A$.
   c) If $A$ is a finite set, then $|[u]_{\rho_B}| = |[u]_{\sigma_B} = |B|$ for every $u \in G$.
   d) If $A$ is finite and $B$ is a subgroup of $\mathcal{A}$, then $|B|$ divides $|A|$.
9. If $B$ is a subgroup of group $\mathcal{A} = (A, \{e, \cdot, ^{-1}\})$, let $xB = \{x \cdot y \mid y \in B\}$ and $Bx = \{y \cdot x \mid y \in B\}$. Prove that $\rho_B = \sigma_B$ if and only if $xB = Bx$ for every $x \in A$; also, show that in this case $\rho_B$ is a congruence of $\mathcal{A}$.
10. Let $\mathcal{A}$ and $\mathcal{B}$ be two algebras of the same type. Prove that $f : A \longrightarrow B$ belongs to $\mathsf{MOR}(\mathcal{A}, \mathcal{B})$ if and only if the set $\{(x, h(x)) \mid x \in A\}$ is a subalgebra of the product algebra $\mathcal{A} \times \mathcal{B}$.

Let $\mathcal{A} = (A, \mathcal{I})$ be an algebra. The set $\mathsf{Pol}_n(\mathcal{A})$ of *n-ary polynomials of the algebra* $\mathcal{A}$ consists of the following functions:
(i) Every projection $p_i : A^n \longrightarrow A$ is an $n$-ary polynomial.
(ii) If $f$ is an $m$-ary operation and $g_1, \ldots, g_m$ are $n$-ary polynomials, then $f(g_1, \ldots, g_n)$ is an $n$-ary polynomial.
The set of polynomials of the algebra $\mathcal{A}$ is the set $\mathsf{Pol}(\mathcal{A}) = \bigcup_{n \in \mathbb{N}} \mathsf{Pol}_n(\mathcal{A})$.

A *k-ary algebraic function* of $\mathcal{A}$ is a function $h : A^k \longrightarrow A$ for which there exists a polynomial $p \in \mathsf{Pol}_n(\mathcal{A})$ and $n - k$ elements $a_{i_1}, \ldots, a_{i_{n-k}}$ of $A$ such that $h(x_1, \ldots, x_k) = p(x_1, \ldots, a_{i_1}, \ldots, a_{i_{n-k}}, \ldots, x_k)$ for $x_1, \ldots, x_k \in A$.

11. Let $\rho$ be a congruence of an algebra $\mathcal{A} = (A, \mathcal{I})$ and let $f \in \mathsf{Pol}_n(\mathcal{A})$. Prove that if $(x_i, y_i) \in \rho$ for $1 \leq i \leq n$, then $(f(x_1, \ldots, x_n), f(y_1, \ldots, y_n)) \in \rho$.

12. Let $\mathcal{A} = (A, \mathcal{I})$ be an algebra and let $S$ be a subset of $A$. Define the sequence of sets $\mathbf{S} = (S_0, S_1, \ldots)$ as $S_0 = S$ and $S_{n+1} = S_n \cup \{f(a_1, \ldots, a_m) \mid f \text{ is an } m\text{-ary operation}, a_1, \ldots, a_m \in S_n\}$.
    a) Prove that the least subalgebra of $\mathcal{A}$ that contains $S$ is $\bigcup_{n \in \mathbb{N}} S_n$.
    b) Prove by induction on $n$ that if $a \in S_n$, then there is a finite subset $U$ of $A$ such that $a$ belongs to the least subalgebra that contains $U$.
13. Let $\mathcal{A} = (A, \mathcal{I})$ be an algebra, $S$ be a subset of $A$, and $a$ be an element in the least subalgebra of $\mathcal{A}$ that contains $S$. Prove that there is a finite subset $T$ of $S$ such that $a$ belongs to the least subalgebra of $\mathcal{A}$ that contains $T$.

Let $\mathfrak{A} = \{\mathcal{A}_i \mid i \in I\}$ be a collection of algebras of the same type indexed by the set $I$, where $\mathcal{A}_i = (A_i, \mathcal{I})$. The *product* of the collection $\mathfrak{A}$ is the algebra $\prod_{i \in I} \mathcal{A}_i = (\prod_{i \in I} A_i, \mathcal{I})$, whose operations are defined componentwise, as follows. If $f \in \mathcal{I}$ is an $n$-ary operation and $t_1, \ldots, t_n \in \prod_{i \in I} A_i$, where $t_k = (t_{ki})_{i \in I}$ for $1 \leq k \leq n$, then $f(t_1, \ldots, t_n) = s$, where $s = (s_i)_{i \in I}$ and $s_i = f(t_{1i}, \ldots, t_{ni})$ for $i \in I$.

14. Let $\mathfrak{A} = \{\mathcal{A}_i \mid i \in I\}$ be a collection of algebras of the same type $\theta$ indexed by the set $I$, where $\mathcal{A}_i = (A_i, \mathcal{I})$, and let $\mathcal{A} = \prod_{i \in I} \mathcal{A}_i$ be their product. Prove that each projection $p_i : \prod_{i \in I} A_i \longrightarrow A_i$ belongs to $\mathsf{MOR}(\mathcal{A}, \mathcal{A}_i)$. Furthermore, prove that if $\mathcal{B}$ is an algebra of type $\theta$ and $h_i \in \mathsf{MOR}(\mathcal{B}, \mathcal{A}_i)$ for every $i \in I$, then there exists a unique morphism $h \in \mathsf{MOR}(\mathcal{B}, \mathcal{A})$ such that $h_i = p_i h$ for every $i \in I$.
15. Let $(L, +, \cdot)$ be an $\mathcal{F}$-linear space. Prove that if $a\mathbf{x} = 0$ for $a \in F$ and $\mathbf{x} \in L$, then either $a = 0$ or $\mathbf{x} = \mathbf{0}$.
16. Prove that a subset $K$ of a linear space is linearly dependent if and only of there is $\mathbf{x} \in K$ that can be expressed as a linear combination of $K - \{\mathbf{x}\}$.
17. Let $K$ be a finite set that spans the $\mathcal{F}$-linear space $(L, +, \cdot)$ and let $H$ be a subset of $L$ that is linearly independent. There exists a basis $B$ such that $H \subseteq B \subseteq K$.
18. Prove that every norm $\nu$ on $\mathbb{R}^n$ that is generated by an inner product satisfies the equality

$$\nu(\mathbf{x} + \mathbf{y})^2 + \nu(\mathbf{x} - \mathbf{y})^2 = 2(\nu(\mathbf{x})^2 + \nu(\mathbf{y})^2)$$

    for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (the *parallelogram equality*).
19. Prove that the norms $\nu_1$ and $\nu_\infty$ on $\mathbb{R}^n$ are not generated by an inner product.
    **Hint:**Use Exercise 18.
20. Let $\nu$ be a norm on $\mathbb{R}^n$ that satisfies the parallelogram equality as introduced in Exercise 18. Prove that the function $p : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ given by

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{4}\left(\nu(\mathbf{x} + \mathbf{y})^2 - \nu(\mathbf{x} - \mathbf{y})^2\right)$$

    is an inner product on $\mathbb{R}^n$.

21. Let $S$ be a finite set $S = \{x_1, \ldots, x_n\}$ and let $*$ be a binary operation on $S$. For $t$ in $S$, define the matrices $\mathbf{L}_t, \mathbf{M}_t \in S^{n\times n}$ as $(\mathbf{L}_t)_{ij} = u$ if $(x_i * t) * x_j = u$ and $(\mathbf{R}_t)_{ij} = v$ if $x_i * (t * x_j) = u$.
Prove that "$*$" is an associative operation on $S$ if and only if for every $t \in S$ we have $\mathbf{L}_t = \mathbf{R}_t$.
22. Let $\mathbf{A} = (a_{ij})$ be an $(m \times n)$-matrix of real numbers. Prove that

$$\max_j \min_i a_{ij} \leq \min_i \max_j a_{ij}$$

(the *minimax inequality*).

**Solution:** Note that $a_{ij_0} \leq \max_j a_{ij}$ for every $i$ and $j_0$, so $\min_i a_{ij_0} \leq \min_i \max_j a_{ij}$, again for every $j_0$. Thus, $\max_j \min_i a_{ij} \leq \min_i \max_j a_{ij}$.

## Bibliographical Comments

This chapter is a limited introduction to algebras and linear spaces. The readers interested in a deeper study of those structures should consult the vast mathematical literature concerning general and universal algebra [32, 14, 52, 113] as well as linear algebra [85, 58, 55, 121].

# 3

# Graphs and Hypergraphs

## 3.1 Introduction

Graphs model relations between elements of sets. The term "graph" is suggested by the fact that these mathematical structures can be graphically represented. We discuss two types of graphs: directed graphs, which are suitable for representing arbitrary binary relations, and undirected graphs that are useful for representing symmetric binary relations. Special attention is paid to trees, a type of graph that plays a prominent role in many data mining tasks.

## 3.2 Basic Notions of Graph Theory

**Definition 3.1.** *An* undirected graph*, or simply a* graph*, is a pair* $\mathcal{G} = (V, E)$*, where* $V$ *is a set of* vertices *or* nodes *and* $E$ *is a collection of two-element sets. If* $\{x, y\} \in E$*, we say that* $e = \{x, y\}$ *is an* edge *of* $\mathcal{G}$ *that joins* $x$ *to* $y$*. The vertices* $x$ *and* $y$ *are the* endpoints *of the edge* $e$*.*

*A graph* $\mathcal{G} = (V, E)$ *is* finite *if both* $V$ *and* $E$ *are finite. The number of vertices* $|V|$ *is referred to as the* order of the graph*.*

If $u$ is an endpoint of an edge $e$, we say that $e$ is incident to $u$. To simplify the notation, we denote an edge $e = \{x, y\}$ by $(x, y)$. If $e = (x, y)$ is an edge, we say that $x$ and $y$ are *adjacent* vertices. If $e$ and $e'$ are two distinct edges in a graph $\mathcal{G}$, then $|e \cap e'| \leq 1$.

Graphs can be drawn by representing each vertex by a point in a plane and each edge $(x, y)$ by an arc joining $x$ and $y$.

*Example 3.2.* Figure 3.1 contains the drawing of the graph $\mathcal{G} = (\{v_i \mid 1 \leq i \leq 8\}, E)$, where

$$E = \{(v_1, v_2), (v_1, v_3), (v_2, v_3), (v_4, v_5), \\ (v_5, v_6), (v_6, v_7), (v_7, v_8), (v_5, v_8)\}.$$

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_3, © Springer-Verlag London Limited 2008

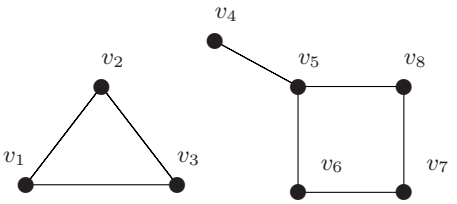



**Fig. 3.1.** Graph $\mathcal{G} = (\{v_i \mid 1 \leq i \leq 8\}, E)$.

An equivalent representation is to use a two-column table $T_{\mathcal{G}}$ in which each row represents an edge $(v_i, v_j)$, where $i < j$. For example, the graph introduced in Example 3.2 can also be represented in tabular form, as shown in Table 3.1.

**Table 3.1.** Tabular representation of a graph.

| First Vertex | Second Vertex |
|---|---|
| $v_1$ | $v_2$ |
| $v_1$ | $v_3$ |
| $v_2$ | $v_3$ |
| $v_4$ | $v_5$ |
| $v_5$ | $v_6$ |
| $v_6$ | $v_7$ |
| $v_7$ | $v_8$ |
| $v_5$ | $v_8$ |

The set of vertices that are adjacent to the vertex $x$ in the graph $\mathcal{G}$ is denoted by $\Gamma_{\mathcal{G}}(x)$, or simply by $\Gamma(x)$ if the graph is clear from the context.

### 3.2.1 Degrees of Vertices

**Definition 3.3.** *Let $\mathcal{G} = (V, E)$ be a graph. The* degree of a vertex $v$ *is the number*

$$\mathsf{d}_{\mathcal{G}}(v) = |\Gamma_{\mathcal{G}}(x)|.$$

When the graph $\mathcal{G}$ is clear from the context, we omit the subscript and simply write $\mathsf{d}(x)$ and $\Gamma(x)$ instead of $\mathsf{d}_{\mathcal{G}}(x)$ and $\Gamma_{\mathcal{G}}(x)$, respectively.

If $\mathsf{d}(v) = 0$, then $v$ is an *isolated vertex*. Note that the degree of a vertex $v$ equals the number of rows of the table in which $v$ occurs and the total number of rows equals the number of edges. For a graph $\mathcal{G} = (V, E)$, we have

$$\sum\{\mathsf{d}(v) \mid v \in V\} = 2|E| \tag{3.1}$$

because when adding the degrees of the vertices of the graph we count the number of edges twice. Since the sum of the degrees is an even number, it follows that a finite graph has an even number of vertices that have odd degrees. Also, for every vertex $v$, we have $\mathsf{d}(v) \leq |V| - 1$.

**Definition 3.4.** *A sequence* $(d_1, \ldots, d_n) \in \mathbf{Seq}_n(\mathbb{N})$ *is a* graphic sequence *if there is a graph* $\mathcal{G} = (\{v_1, \ldots, v_n\}, E)$ *such that* $\mathsf{d}(v_i) = d_i$ *for* $1 \leq i \leq n$.

Clearly, not every sequence of natural numbers is graphic since we must have $d_i \leq n - 1$ and $\sum_{i=1}^{n} d_i$ must be an even number. For example, the sequence $(5, 5, 4, 3, 3, 2, 1)$ is not graphic since the sum of its components is not even. A characterization of graphic sequences obtained in [66, 60] is given next.

**Theorem 3.5 (The Havel-Hakimi Theorem).** *Let* $\boldsymbol{d} = (d_1, \ldots, d_n)$ *be a sequence of natural numbers such that* $d_1 \geq d_2 \geq \cdots \geq d_n$, $n \geq 2$, $d_1 \geq 1$, *and* $d_i \leq n-1$ *for* $1 \leq i \leq n$. *The sequence* $\boldsymbol{d}$ *is graphic if and only if the sequence*

$$\boldsymbol{e} = (d_2 - 1, d_3 - 1, \ldots, d_{d_1+1} - 1, d_{d_1+2}, \ldots, d_n)$$

*is graphic.*

*Proof.* Suppose that $\mathbf{d} = (d_1, \ldots, d_n)$ is a graphic sequence, and let $\mathcal{G} = (\{v_1, \ldots, v_n\}, E)$ be a graph having $\mathbf{d}$ as the sequence of degrees of its vertices.

If there exists a vertex $v_1$ of degree $d_1$ that is adjacent with vertices having degrees $d_2, d_3, \ldots, d_{d_1+1}$, then the graph $\mathcal{G}'$ obtained from $\mathcal{G}$ by removing the vertex $v_1$ and the edges having $v_1$ as an endpoint has $\mathbf{e}$ as its degree sequence, so $\mathbf{e}$ is graphic.

If no such vertex exists, then there are vertices $v_i, v_j$ such that $i < j$ (and thus $d_i \geq d_j$), such that $(v_1, v_j)$ is an edge but $(v_1, v_i)$ is not. Since $d_i \geq d_j$ there exists a vertex $v_k$ such that $v_k$ is adjacent to $v_i$ but not to $v_j$.

Now let $\hat{\mathcal{G}}$ be the graph obtained from $\mathcal{G}$ by removing the edges $(v_1, v_j)$ and $(v_i, v_k)$ shown in Figure 3.2(a) and adding the edges $(v_1, v_i)$ and $(v_j, v_k)$ shown in Figure 3.2(b). Observe that the degree sequence of $\hat{\mathcal{G}}$ remains the same but the sum of the degrees of the vertices adjacent to $v_1$ increases. This process may be repeated only a finite number of times before ending with a graph that belongs to the first case.

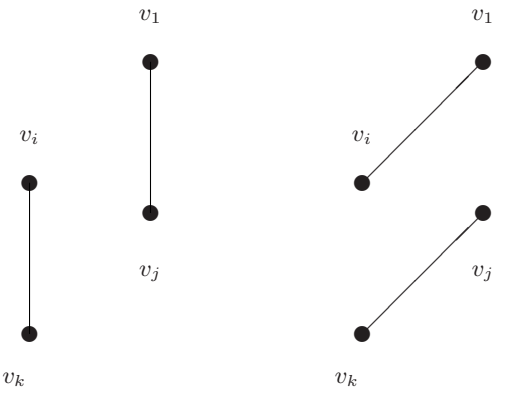



**Fig. 3.2.** Construction of graph $\hat{\mathcal{G}}$.

Conversely, suppose that $\mathbf{e}$ is a graphic sequence, and let $\mathcal{G}_1$ be a graph that has $\mathbf{e}$ as the sequence of vertex degrees. Let $\mathcal{G}_2$ be a graph obtained from $\mathcal{G}_1$ by adding a new vertex $v$ adjacent to vertices of degrees $d_2-1, d_3-1, \ldots, d_{d_1+1}-1$. Clearly, the new vertex has degree $d_1$ and the degree sequence of the new graph is precisely $\mathbf{d}$. □

*Example 3.6.* Let us determine if the sequence $(5, 4, 4, 3, 3, 3, 3, 1)$ is a graphic sequence. Note that the sum of its components is an even number. The sequence derived from it by applying the transformation of Theorem 3.5 is $(3, 3, 2, 2, 2, 3, 1)$. Rearranging the sequence in nonincreasing order, we have the same question for the sequence $(3, 3, 3, 2, 2, 2, 1)$. A new transformation yields the sequence $(2, 2, 1, 2, 2, 1)$. Placing the components of this sequence in increasing order yields $(2, 2, 2, 2, 1, 1)$. A new transformation produces the shorter sequence $(1, 1, 2, 1, 1)$. The new resulting sequence $(2, 1, 1, 1, 1)$ can be easily seen to be the degree sequence of the graph shown in Figure 3.3(a). We show the degree of each vertex.

The process is summarized in Table 3.2.

**Table 3.2.** Degree sequences.

| $d_1$ | Sequence |
|---|---|
| | $(5, 4, 4, 3, 3, 3, 3, 1)$ |
| 5 | $(3, 3, 3, 2, 2, 2, 1)$ |
| 3 | $(2, 2, 2, 2, 1, 1)$ |
| 2 | $(2, 1, 1, 1, 1)$ |

Starting from the graph having degree sequence $(2, 1, 1, 1, 1)$, we add a new vertex and two edges linking this vertex to two vertices of degree 1 to obtain the graph of Figure 3.3(b), which has the degree sequence $(2, 2, 2, 2, 1, 1)$.

In the next step, a new vertex is added that is linked by three edges to vertices of degrees $2, 2$, and 1. The resulting graph shown in Figure 3.3(c) has the degree sequence $(3, 3, 3, 2, 2, 2, 1)$. Finally, a vertex of degree 5 is added to produce the graph shown in Figure 3.3(d), which has the desired degree sequence.

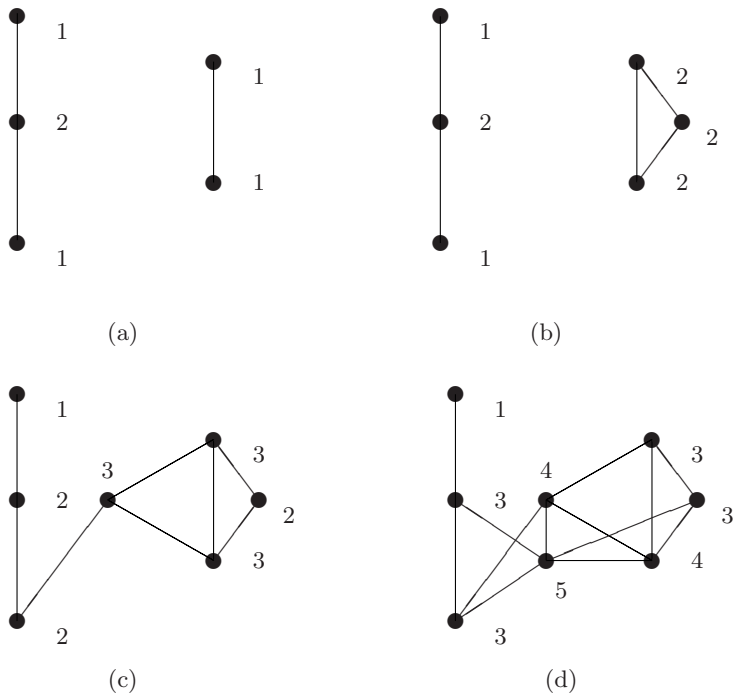



**Fig. 3.3.** Construction of a graph with a prescribed degree sequence.

A graph $\mathcal{G}$ is *k-regular* if all vertices have the same degree. If $\mathcal{G}$ is $k$-regular for some $k$, then we say that the graph is *regular*.

**Definition 3.7.** *A* bipartite graph *is a graph $\mathcal{G} = (V, E)$ such that there is a two-block partition $\pi = \{V_1, V_2\}$ for which $E \subseteq V_1 \times V_2$. If $E = V_1 \times V_2$, then we say that $\mathcal{G}$ is* bipartite complete.

In other words, $\mathcal{G}$ is bipartite if every edge of the graph has its endpoints in two distinct classes.

A bipartite complete graph, where $|V_1| = p$ and $|V_2| = q$ is denoted by $\mathcal{K}_{p,q}$.

*Example 3.8.* The bipartite graph $\mathcal{K}_{3,3}$ is shown in Figure 3.4.

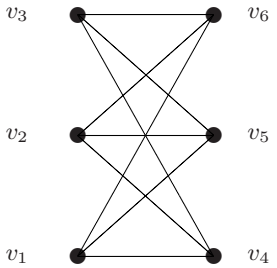



**Fig. 3.4.** Bipartite graph $\mathcal{K}_{3,3}$.

### 3.2.2 Graph Representations

Finite graphs are often represented using matrices.

**Definition 3.9.** *Let $\mathcal{G} = (V, E)$ be a finite graph, where $V = \{v_1, \ldots, v_m\}$ and $E = \{e_1, \ldots, e_n\}$.*

*The* incidence matrix *of $\mathcal{G}$ is the $m \times n$ matrix $\boldsymbol{I}_{\mathcal{G}} = (i_{pr})$ given by*

$$i_{pr} = \begin{cases} 1 & \text{if } v_p \text{ is incident to } e_r, \\ 0 & \text{otherwise,} \end{cases}$$

*for $1 \leq p \leq m$ and $1 \leq r \leq n$.*

*The* adjacency matrix *of $\mathcal{G}$ is the $m \times m$ matrix $\boldsymbol{A}_{\mathcal{G}} = (a_{pq})$ given by*

$$a_{pq} = \begin{cases} 1 & \text{if } v_p \text{ is adjacent to } v_q, \\ 0 & \text{otherwise,} \end{cases}$$

*for $1 \leq p, q \leq m$.*

*Example 3.10.* Let $\mathcal{G}$ be the graph shown in Figure 3.5. Its incidence matrix is the $6 \times 7$ matrix

$$\mathbf{I}_{\mathcal{G}} = \begin{pmatrix} 1\,0\,0\,0\,0\,0\,0 \\ 1\,0\,0\,0\,0\,0\,0 \\ 0\,1\,1\,0\,0\,1\,0 \\ 0\,0\,1\,1\,0\,0\,1 \\ 0\,1\,0\,1\,1\,0\,0 \\ 0\,0\,0\,0\,1\,1\,1 \end{pmatrix}.$$

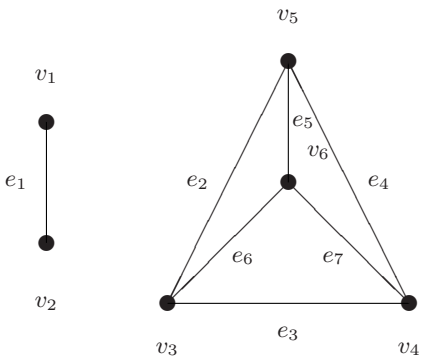



**Fig. 3.5.** Graph $\mathcal{G} = (\{v_1, \ldots, v_5\}, \{e_1, \ldots, e_7\})$.

The adjacency matrix is the $6 \times 6$-matrix

$$\mathbf{A} = \begin{pmatrix} 0&1&0&0&0&0 \\ 1&0&0&0&0&0 \\ 0&0&0&1&1&1 \\ 0&0&1&0&1&1 \\ 0&0&1&1&0&1 \\ 0&0&1&1&1&0 \end{pmatrix}.$$

### 3.2.3 Paths

**Definition 3.11.** *A* path of length $n$ *in a graph* $\mathcal{G} = (V, E)$ *is a sequence of vertices* $\boldsymbol{p} = (v_0, \ldots, v_n)$ *such that* $(v_i, v_{i+1})$ *is an edge of* $\mathcal{G}$ *for* $0 \leq i \leq n-1$. *The vertices* $v_0$ *and* $v_n$ *are the* endpoints *of* $\boldsymbol{p}$, *and we say that the path* $\boldsymbol{p}$ connects the vertices $v_0$ and $v_n$. *The length of the path* $\boldsymbol{p}$ *will be denoted by* $\ell(p)$.

*A path is* simple *if all vertices of the path are distinct.*

*A path* $\boldsymbol{p} = (v_0, \ldots, v_n)$ *is a* cycle *if* $n \geq 3$ *and* $v_0 = v_n$. *A cycle is* simple *if all vertices are distinct with the exception of the first and the last. A cycle of length* 3 *is called a* triangle.

*A graph with no cycles is said to be* acyclic.

For every vertex $v$ of a graph $\mathcal{G} = (V, E)$, there is a unique path $(v)$ of length 0 that joins $v$ to itself.

**Definition 3.12.** *Let* $\mathcal{G} = (V, E)$ *be a graph and let* $x, y \in V$ *be two vertices. The* distance $d(x, y)$ between $x$ and $y$ *is the length of the shortest path that has* $x$ *and* $y$ *as its endpoints. If no such path exists, then we define* $d(x, y) = \infty$.

*Example 3.13.* The distances between the vertices of the graph shown in Figure 3.1 are given in the following table.

| $d$ | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $v_7$ | $v_8$ |
|---|---|---|---|---|---|---|---|---|
| $v_1$ | 0 | 1 | 1 | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| $v_2$ | 1 | 0 | 1 | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| $v_3$ | 1 | 1 | 0 | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| $v_4$ | $\infty$ | $\infty$ | $\infty$ | 0 | 1 | 2 | 3 | 2 |
| $v_5$ | $\infty$ | $\infty$ | $\infty$ | 1 | 0 | 1 | 2 | 1 |
| $v_6$ | $\infty$ | $\infty$ | $\infty$ | 2 | 1 | 0 | 1 | 2 |
| $v_7$ | $\infty$ | $\infty$ | $\infty$ | 3 | 2 | 1 | 0 | 1 |
| $v_8$ | $\infty$ | $\infty$ | $\infty$ | 2 | 1 | 2 | 1 | 0 |

### 3.2.4 Directed Graphs

Directed graphs differ from graphs in that every edge in such a graph has an orientation. The formal definition that follows captures this aspect of directed graphs by defining an edge as an ordered pair of vertices rather than a two-element set.

**Definition 3.14.** *A* directed graph *(or, for short, a* digraph *) is a pair* $\mathcal{G} = (V, E)$*, where* $V$ *is a set of* vertices *or* nodes *and* $E \subseteq V \times V$ *is the set of* edges.

*A digraph* $\mathcal{G} = (V, E)$ *is* finite *if both* $V$ *and* $E$ *are finite.*

If $e = (u, v) \in E$, we refer to $u$ as *the source of the edge* $e$ and to $v$ as *the destination of* $e$. The source and the destination of an edge $e$ are denoted by $source(e)$ and $dest(e)$, respectively. Thus, we have the mappings $source : E \longrightarrow V$ and $dest : E \longrightarrow V$, which allow us to define for every subset $U$ of the set of vertices the sets

$$out(U) = \{source^{-1}(U) - dest^{-1}(U)\},$$
$$in(U) = \{dest^{-1}(U) - source^{-1}(U)\}.$$

In other words, $out(U)$ is the set of edges that originate in $U$ without ending in this set and $in(U)$ is the set of edges that end in $U$ without originating in $U$.

**Definition 3.15.** *Let* $\mathcal{G} = (V, E)$ *be a digraph. The* in-degree of a vertex $v$ *is the number*

$$\mathsf{d}_o(v) = |\{e \in E \mid source(e) = v\}|,$$

*and the* out-degree of a vertex $v$ *is the number*

$$\mathsf{d}_i(v) = |\{e \in E \mid dest(e) = v\}|.$$

Clearly, we have

$$\sum_{v \in V} \mathsf{d}_o(v) = \sum_{v \in V} \mathsf{d}_i(v) = |E|.$$

The notion of a path for digraphs is similar to the notion of a path for graphs.

**Definition 3.16.** *Let $\mathcal{G} = (V, E)$ be a digraph. A* path *in $\mathcal{G}$ is a sequence of vertices $\boldsymbol{p} = (v_0, \ldots, v_n)$ such that $(v_i, v_{i+1}) \in E$ for every $i$, $0 \leq i \leq n-1$. The number $n$ is the* length *of $\boldsymbol{p}$. We refer to $\boldsymbol{p}$ as a* path that joins $v_0$ to $v_{n-1}$.

*If all vertices of the sequence $(v_0, \ldots, v_{n-1}, v_n)$ are distinct, with the possible exception of $v_0$ and $v_n$, then $\boldsymbol{p}$ is a* simple path.

*If $v_0 = v_n$, then $\boldsymbol{p}$ is a* cycle. *A directed graph with no cycles is said to be* acyclic.

Note that a path $\mathbf{p}$ may have length 0; in this case, $\mathbf{p}$ is the null sequence of edges and the sequence of vertices of $\mathbf{p}$ consists of a single vertex.

If there is a path $\mathbf{p}$ from $u$ to $v$ in a directed acyclic graph, then there is no path from $v$ to $u$, where $u$ and $v$ are two distinct nodes of $\mathcal{G}$, because otherwise we would have a cycle.

**Definition 3.17.** *Let $\mathcal{G} = (V, E)$ be an acyclic digraph and let $u, v \in V$. The vertex $u$ is an* ancestor *of $v$, and $v$ is a* descendant *of $u$ if there is a path $\boldsymbol{p}$ from $u$ to $v$.*

Note that every vertex is both an ancestor and a descendant of itself due to the existence of paths of length 0. If $u$ is an ancestor of $v$ and $u \neq v$, then we say that $u$ is a *proper ancestor* of $v$. Similarly, if $v$ is a descendant of $u$ and $u \neq v$, we refer to $v$ as a *proper descendant* of $u$.

**Definition 3.18.** *A* forest *is an acyclic digraph $\mathcal{G} = (V, E)$ such that $\mathsf{d}_i(v) \leq 1$ for every vertex $v$.*

**Theorem 3.19.** *Let $V_0$ be the set of vertices of a finite forest $\mathcal{G} = (V, E)$ having in-degree 0. For every vertex $v \in V - V_0$, there exists a unique vertex $v_0 \in V_0$ and a unique path that joins $v_0$ to $v$.*

*Proof.* Since $v \in V - V_0$, we have $\mathsf{d}_i(v) = 1$ and there exists at least one edge whose destination is $v$. Let $\mathbf{p} = (v_0, v_1, \ldots, v_{n-1})$ be a maximal path whose destination is $v$ (so $v_{n-1} = v$). We have $\mathsf{d}_i(v_0) = 0$. Indeed, if this is not the case, then there is a vertex $v'$ such that an edge $(v', v)$ exists in $E$ and $v'$ is distinct from every vertex of $\mathbf{p}$ because otherwise $\mathcal{G}$ would not be acyclic. This implies the existence of a path $\mathbf{p}' = (v', v_0, v_1, \ldots, v_{n-1})$, which contradicts the maximality of $\mathbf{p}$.

The path that joins a vertex of in-degree 0 to $v$ is unique. Indeed, suppose that $\mathbf{q}$ is another path in $\mathcal{G}$ that joins a vertex of in-degree 0 to $v$. Let $u$ be the first vertex having a positive in-degree that is common to both paths. The predecessors of $u$ on $\mathbf{p}$ and $\mathbf{q}$ must be distinct, which implies that $\mathsf{d}_i(u) > 1$. This contradicts the fact that $\mathcal{G}$ is a forest. The uniqueness of the path implies also the uniqueness of the source. □

**Theorem 3.20.** *Let $\mathcal{G} = (V, E)$ be a graph. The relation $\gamma_{\mathcal{G}}$ on $V$ that consists of all pairs of vertices $(x, y)$ such that there is a path that joins $x$ to $y$ is an equivalence on $V$.*

*Proof.* The reflexivity of $\gamma_{\mathcal{G}}$ follows from the fact that there is a path of length 0 that joins any vertex $x$ to itself.

If a path $\mathbf{p} = (v_0, \ldots, v_n)$ joins $x$ to $y$ (which means that $x = v_0$ and $y = v_n$), then the path $\mathbf{q} = (v_n, \ldots, v_0)$ joins $y$ to $x$. Therefore, $\gamma_{\mathcal{G}}$ is symmetric.

Finally, suppose that $(x, y) \in \gamma_{\mathcal{G}}$ and $(y, z) \in \gamma_{\mathcal{G}}$. There is a path $\mathbf{p} = (v_0, \ldots, v_n)$ with $x = v_0$ and $y = v_n$ and a path $\mathbf{q} = (v'_0, \ldots, v'_m)$ such that $y = v'_0$ and $z = v'_m$. The path $\mathbf{r} = (v_0, \ldots, v_n = v'_0, \ldots, v'_m)$ joins $x$ to $z$, so $(x, z) \in \gamma_{\mathcal{G}}$. Thus, $\gamma_{\mathcal{G}}$ is transitive, so it is an equivalence. □

**Definition 3.21.** *Let $\mathcal{G} = (V, E)$ be a graph. The* connected components *of $\mathcal{G}$ are the equivalence classes of the relation $\gamma_{\mathcal{G}}$.*

*A graph is* connected *if it has only one connected component.*

*Example 3.22.* The sequences $(v_4, v_5, v_8, v_7)$ and $(v_4, v_5, v_6, v_7)$ are both paths of length 3 in the graph shown in Figure 3.1.

The connected components of this graph are

$$\{v_1, v_2, v_3\} \text{ and } \{v_4, v_5, v_6, v_7, v_8\}.$$

**Definition 3.23.** *A* subgraph *of a graph $\mathcal{G} = (V, E)$ is a graph $\mathcal{G}' = (V', E')$, where $V' \subseteq V$ and $E' \subseteq E$.*

*The* subgraph of $\mathcal{G}$ induced by a set of vertices $U$ *is the subgraph $\mathcal{G}_U = (U, E_U)$, where $E_U = \{e \in E \mid e = (u, u') \text{ and } u, u' \in U\}$.*

*A* spanning subgraph *of a graph $\mathcal{G} = (V, E)$ is a subgraph of the form $\mathcal{G}' = (V, E')$; that is, a subgraph that has the same set of vertices as $\mathcal{G}$.*

*Example 3.24.* Consider the graph $\mathcal{G}$ shown in Figure 3.1. In Figures 3.6(a)-(d), we show the subgraphs of the graph $\mathcal{G}$ induced by the sets

$$\{v_4, v_5, v_8\}, \{v_4, v_5, v_6, v_8\}, \{v_4, v_5, v_7\}, \{v_5, v_6, v_7, v_9\},$$

respectively.

*Example 3.25.* The graph shown in Figure 3.7 is a spanning subgraph of the graph defined in Example 3.2.

**Theorem 3.26.** *If $\mathcal{G} = (V, E)$ is a connected graph, then $|E| \geq |V| - 1$.*

*Proof.* We prove the statement by induction on $|E|$. If $|E| = 0$, then $|V| \leq 1$ because $\mathcal{G}$ is connected and the inequality is clearly satisfied.

Suppose that the inequality holds for graphs having fewer than $n$ edges, and let $\mathcal{G} = (V, E)$ be a connected graph with $|E| = n$. Let $e = (x, y)$ be an arbitrary edge and let $\mathcal{G}' = (V, E - \{e\})$ be the graph obtained by removing the edge $e$. The graph $\mathcal{G}'$ may have one or two connected components, so we need to consider the following cases:

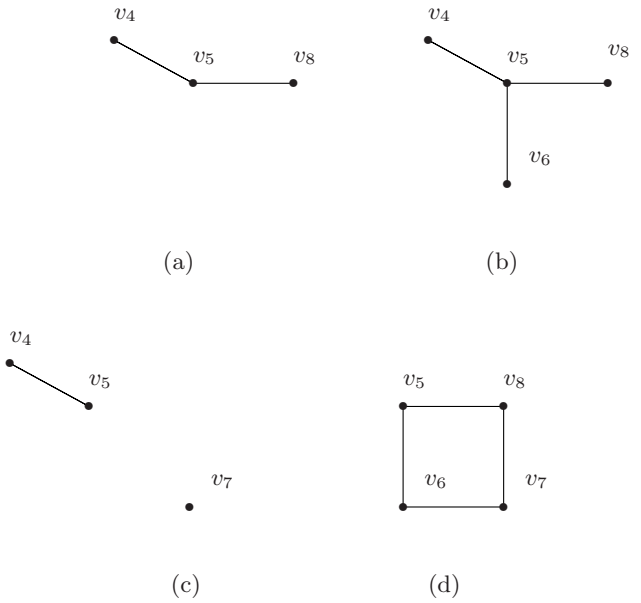



**Fig. 3.6.** Subgraphs of the graph $\mathcal{G}$.

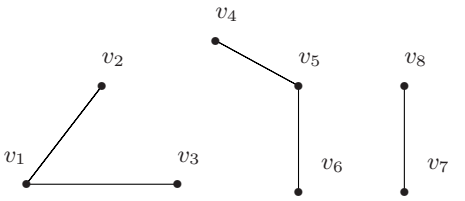



**Fig. 3.7.** Spanning subgraph of the graph defined in Example 3.2.

1. If $\mathcal{G}'$ is connected, then, by the inductive hypothesis, we have $|E'| \geq |V|-1$, which implies $|E| = |E'| + 1 \geq |V| - 1$.
2. If $\mathcal{G}'$ contains two connected components $V_0$ and $V_1$, let $E_0$ and $E_1$ be the set of edges whose endpoints belong to $V_0$ and $V_1$, respectively. By the inductive hypothesis, $|E_0| \geq |V_0| - 1$ and $|E_1| \geq |V_1| - 1$. This implies

$$|E| = |E_0| + |E_1| + 1 \geq |V_0| + |V_1| - 1 = |V| - 1.$$

This concludes the argument. □

**Corollary 3.27.** *Let $\mathcal{G} = (V, E)$ be a graph that has $k$ connected components. We have $|E| \leq |V| - k$.*

*Proof.* Let $V_i$ and $E_i$ be the set of edges of the $i^{\text{th}}$ connected component of $\mathcal{G}$, where $1 \leq i \leq k$. It is clear that

$$V = \bigcup_{i=1}^{k} V_i \text{ and } E = \bigcup_{i=1}^{k} E_i.$$

Since the sets $V_1, \ldots, V_k$ form a partition of $V$ and the sets $E_1, \ldots, E_k$ form a partition of $E$, we have

$$|E| = \sum_{i=1}^{k} |E|_i \leq \sum_{i=1}^{k} |V|_i - k = |V| - k,$$

which is the desired inequality. □

**Definition 3.28.** *A graph $\mathcal{G} = (V, E)$ is* complete*, if for every $u, v \in E$ such that $u \neq v$, we have $(u, v) \in E$.*

The complete graph $\mathcal{G} = (\{1, \ldots, n\}, E)$ will be denoted by $\mathcal{K}_n$.

*Example 3.29.* The graph shown in Figure 3.8 is complete.

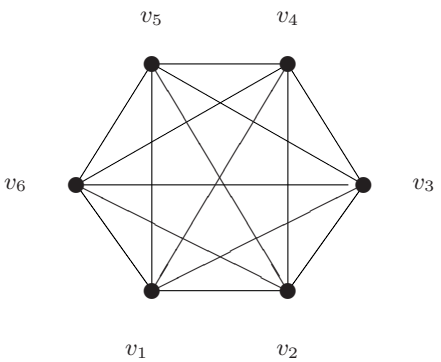



**Fig. 3.8.** Complete graph.

A subset $U$ of the set of vertices of a graph $\mathcal{G} = (V, E)$ is *complete* if the subgraph induced by it is complete.

A set of vertices $W$ is a *clique in* $\mathcal{G}$ if it is maximally complete. In other words, $W$ is a clique if the graph induced by $W$ is complete and there is no set of vertices $Z$ such that $W \subset Z$ and $Z$ is complete.

*Example 3.30.* The cliques of the graph shown in Figure 3.9 are $\{v_1, v_2, v_3, v_4\}$ and $\{v_3, v_5, v_6\}$.

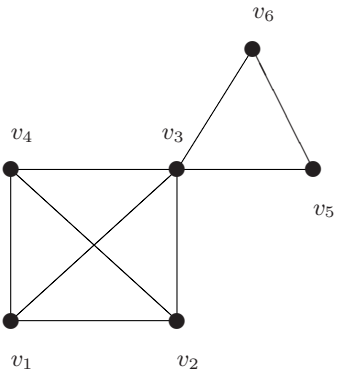



**Fig. 3.9.** Graph and its two cliques.

**Definition 3.31.** *Let* $\mathcal{G}_i = (V_i, E_i)$ *be two graphs, where* $i \in \{1, 2\}$. *The graphs* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *are* isomorphic *if there exists a bijection* $f : V_1 \longrightarrow V_2$ *such that* $(f(u), f(v)) \in E'$ *if and only if* $(u, v) \in E$. *The mapping* $f$ *in this case is called a* graph isomorphism. *If* $\mathcal{G}_1 = \mathcal{G}_2$, *then we say that* $f$ *is a graph* automorphism *of* $\mathcal{G}_1$.

Two isomorphic graphs can be represented by drawings that differ only with respect to the labels of the vertices.

*Example 3.32.* The graphs

$$\mathcal{G}_1 = (\{v_1, v_2, v_3, v_4, v_5\}, E_1) \text{ and } \mathcal{G}_2 = (\{u_1, u_2, u_3, u_4, u_5\}, E_2),$$

shown in Figures 3.10(a) and (b), respectively, are isomorphic. Indeed, the function $f : \{v_1, v_2, v_3, v_4, v_5\} \longrightarrow \{u_1, u_2, u_3, u_4, u_5\}$ defined by

$$f(v_1) = u_1, f(v_2) = u_3, f(v_3) = u_5, f(v_4) = u_2, f(v_5) = u_4$$

can be easily seen to be a graph isomorphism. On the other hand, both graphs shown in Figure 3.11 have six vertices but cannot be isomorphic. Indeed, the first graph consists of one connected component, while the second has two connected components ($\{u_1, u_3, u_5\}$ and $\{u_2, u_4, u_6\}$).

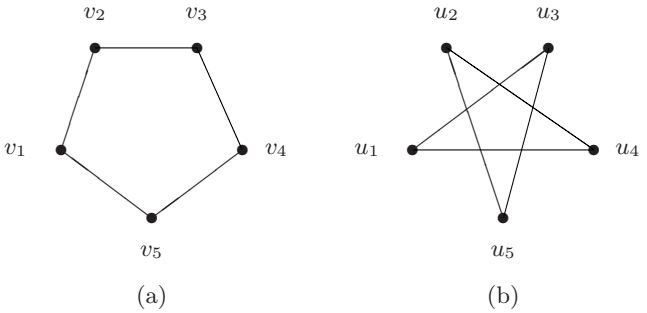



**Fig. 3.10.** Two isomorphic graphs.

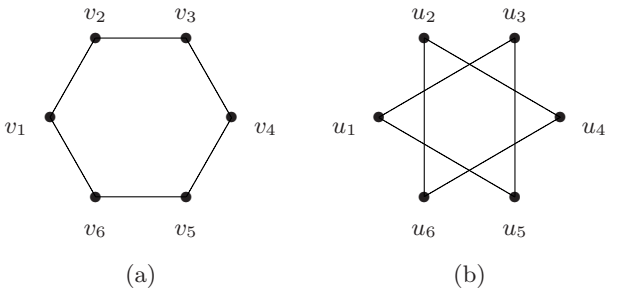



**Fig. 3.11.** Two graphs that are not isomorphic.

If two graphs are isomorphic, they have the same degree sequences. The inverse is not true; indeed, the graphs shown in Figure 3.11 have the same degree sequence $\mathbf{d} = (2, 2, 2, 2, 2, 2)$ but are not isomorphic.

In general, an *invariant* of graphs is a set of numbers that is the same for two isomorphic graphs. Thus, the degree sequence is an invariant of graphs.

## 3.3 Trees

Trees are graphs of special interest to data mining due to the presence of tree-structured data in areas such as Web and text mining and computational biology.

**Definition 3.33.** *A* tree *is a graph* $\mathcal{G} = (V, E)$ *that is both connected and acyclic. A* forest *is a graph* $\mathcal{G} = (V, E)$ *whose connected components are trees.*

*Example 3.34.* The graph shown in Figure 3.12 is a tree having

$$V = \{t, u, v, w, x, y, z\}$$

as its set of vertices and

$$E = \{(t, v), (v, w), (y, u), (y, v), (x, y), (x, z)\}$$

as its set of edges.

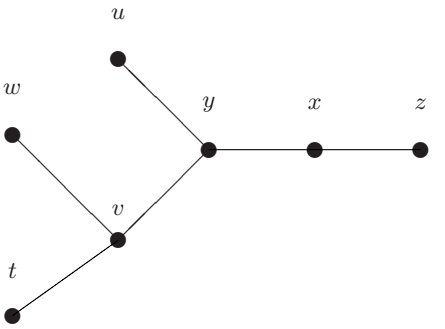



**Fig. 3.12.** Tree having $V = \{t, u, v, w, x, y, z\}$ as its set of vertices.

Next, we give several equivalent characterizations of trees.

**Theorem 3.35.** *Let $\mathcal{G} = (V, E)$ be a graph. The following statements are equivalent:*

(i) *$\mathcal{G}$ is a tree.*
(ii) *Any two vertices $x, y \in V$ are connected by a unique simple path.*
(iii) *$\mathcal{G}$ is minimally connected; in other words, if an edge $e$ is removed from $E$, the resulting graph $\mathcal{G}' = (E, V - \{e\})$ is not connected.*
(iv) *$\mathcal{G}$ is connected, and $|E| = |V| - 1$.*
(v) *$\mathcal{G}$ is an acyclic graph, and $|E| = |V| - 1$.*

*Proof.* (i) implies (ii): Let $\mathcal{G}$ be a graph that is connected and acyclic and let $u$ and $v$ be two vertices of the graph. If $\mathbf{p}$ and $\mathbf{q}$ are two distinct simple paths that connect $u$ to $v$, then $\mathbf{pq}$ is a cycle in $\mathcal{G}$, which contradicts its acyclicity. Therefore, (ii) follows.

(ii) implies (iii): Suppose that any two vertices of $\mathcal{G}$ are connected by a unique simple path. If $e = (u, v)$ is an edge in $\mathcal{G}$, then $(u, v)$ is a simple path in $\mathcal{G}$ and therefore it must be the unique simple path connecting $u$ and $v$. If $e$ is removed, then it is impossible to reach $v$ from $u$, and this contradicts the connectivity of $\mathcal{G}$.

(iii) implies (iv): The argument is by induction on $|V|$. If $|V| = 1$, there is no edge, so the equality is satisfied. Suppose that the statement holds for graphs with fewer than $n$ vertices and that $|V| = n$. Choose an edge $(u, v) \in E$. If the edge $(u, v)$ is removed, then the graph separates into two connected

components $\mathcal{G}_0 = (V_0, E_0)$ and $\mathcal{G}_1 = (V_1, E_1)$ with fewer vertices because $\mathcal{G}$ is minimally connected. By the inductive hypothesis, $|E_0| = |V_0| - 1$ and $|E_1| = |V_1| - 1$. Therefore, $|E| = |E_0| + |E_1| + 1 = |V_0| - 1 + |V_1| - 1 + 1 = |V| - 1$.

(iv) implies (v): Let $\mathcal{G}$ be a connected graph such that $|E| = |V| - 1$. Suppose that $\mathcal{G}$ has a simple cycle $\mathbf{c} = (v_1, \ldots, v_p, v_1)$. Let $\mathcal{G}_0 = (V_0, E_0)$ be the subgraph of $\mathcal{G}$ that consists of the cycle. Clearly, $\mathcal{G}_0$ contains $p$ vertices and an equal number of edges.

Let $U = V - \{v_1, \ldots, v_p\}$. Since $\mathcal{G}$ is connected, there is a vertex $u_1 \in U$ and a vertex $v$ of the cycle $\mathbf{c}$ such that an edge $(u_1, v)$ exists in the graph $\mathcal{G}$. Let $\mathcal{G}_1 = (V_1, E_1)$, where $V_1 = V_0 \cup \{u_1\}$ and $E_1 = E_0 \cup \{(u_1, v)\}$. It is clear that $|V_1| = |E_1|$. If $V - V_1 \neq \emptyset$, there exists a vertex $v_2 \in V - V_2$ and an edge $(v_2, w)$, where $w \in V_1$. This yields the graph $\mathcal{G}_2 = (V_1 \cup \{v_2\}, E_1 \cup \{(v_2, w)\})$, which, again has an equal number of vertices and edges. The process may continue until we exhaust all vertices. Thus, we have a sequence $\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_m$ of subgraphs of $\mathcal{G}$, where $\mathcal{G}_m = (V_m, E_m)$, $|E_m| = |V_m|$ and $V_m = V$. Since $\mathcal{G}_m$ is a subgraph of $\mathcal{G}$, we have $|V| = |V_m| = |E_m| \leq |E|$, which contradicts the fact that $|E| = |V| - 1$. Therefore, $\mathcal{G}$ is acyclic.

(v) implies (i): Let $\mathcal{G} = (V, E)$ be an acyclic graph such that $|E| = |V| - 1$. Suppose that $\mathcal{G}$ has $k$ connected components, $V_1, \ldots, V_k$, and let $E_i$ be the set of edges that connect vertices that belong to the set $V_i$. Note that the graphs $\mathcal{G}_i = (V_i, E_i)$ are both connected and acyclic so they are trees. We have $|E| = \sum_{i=1}^{n} |E_i|$ and $|V| = \sum_{i=1}^{n} |V_i|$. Therefore, $|E| = \sum_{i=1}^{n} |E_i| = \sum_{i=1}^{n} |V_i| - k = |V| - k$. Since $|E| = |V| - 1$ it follows that $k = 1$, so $\mathcal{G} = (V, E)$ is connected. This implies that $\mathcal{G}$ is a tree.

This concludes our argument. □

**Corollary 3.36.** *The graph $\mathcal{G} = (V, E)$ is a tree if and only if it is maximally acyclic; in other words, if an edge $e$ is added to $E$, the resulting graph $\mathcal{G}' = (E, V \cup \{e\})$ contains a cycle.*

*Proof.* Let $\mathcal{G}$ be a tree. If we add an edge $e = (u, v)$ to the $E$, then, since $u$ and $v$ are already connected by a path, we create a cycle. Thus, $\mathcal{G}$ is maximally acyclic.

Conversely, suppose that $\mathcal{G}$ is maximally acyclic. For every pair of vertices $u$ and $v$ in $\mathcal{G}$, two cases may occur:

1. there is an edge $(u, v)$ in $\mathcal{G}$ or
2. there is no edge $(u, v)$ in $\mathcal{G}$.

In the second case, adding the edge $(u, v)$ creates a cycle, which means that there is a path in $\mathcal{G}$ that connects $u$ to $v$. Therefore, in either case, there is a path connecting $u$ to $v$, so $\mathcal{G}$ is a connected graph and therefore a tree. □

**Corollary 3.37.** *If $\mathcal{G} = (V, E)$ is a connected graph, then $\mathcal{G}$ contains a subgraph that is a tree that has $V$ as its set of vertices.*

*Proof.* Define the graph $\mathcal{T} = (V, E')$ as a minimally connected subgraph having the set $V$ as its set of vertices. It is immediate that $\mathcal{T}$ is a tree. □

We shall refer to a tree $\mathcal{T}$ whose existence was shown in Corollary 3.37 as a *spanning tree* of $\mathcal{G}$.

**Corollary 3.38.** *If $\mathcal{G}$ is an acyclic graph, then $\mathcal{G}$ contains $|V| - |E|$ connected components.*

*Proof.* This statement follows immediately from the proof of Theorem 3.35. □

**Definition 3.39.** *A* rooted tree *is a pair $(\mathcal{T}, v_0)$, where $\mathcal{T} = (V, E)$ is a tree and $v_0$ is a vertex of $\mathcal{T}$ called the* root *of $\mathcal{R}$.*

If $(\mathcal{T}, v_0)$ is a rooted tree and $v$ is an arbitrary vertex of $\mathcal{T}$ there is a unique path that joins $v_0$ to $v$. The *height* of $v$ is the length of this path, denoted by $\mathsf{height}(v)$.

The number $\max\{\mathsf{height}(v) \mid v \in V\}$ is the *height* of the rooted tree $(\mathcal{T}, v_0)$; this number is denoted by $\mathsf{height}(\mathcal{T}, v_0)$.

Rooted trees are generally drawn with the root at the top of the picture; if $(u, v)$ is an edge and $\mathsf{height}(u) = \mathsf{height}(v) + 1$, then $u$ is drawn above $v$.

*Example 3.40.* Let $(\mathcal{T}, v_0)$ be the rooted tree shown in Figure 3.13. The heights of the vertices are shown in the following table:

| $v$ | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $v_7$ | $v_8$ |
|---|---|---|---|---|---|---|---|---|
| $\mathsf{height}(v)$ | 1 | 2 | 3 | 1 | 2 | 2 | 3 | 3 |

A rooted tree $(\mathcal{T}, v_0)$ may be regarded as a directed graph. Note that if $(u, v)$ is an edge in a rooted tree, the heights of $u$ and $v$ differ by 1. If $\mathsf{height}(u) = \mathsf{height}(v) + 1$, then we say that $u$ is an *immediate descendant of* $v$ and that $v$ is an immediate ascendant of $u$. The unoriented edge $\{u, v\}$ can now be replaced by the oriented edge $(u, v)$.

In a rooted tree, vertices can be partitioned into sets of nodes named *levels*. Each level $L_i$ consists of those nodes whose height in the tree equals $i$. In a rooted tree $(\mathcal{T}, v_0)$ of height $h$ there exist $h + 1$ levels, $L_0, \cdots, L_h$.

*Example 3.41.* The directed graph that corresponds to the rooted tree from Figure 3.13 is shown in Figure 3.14.

The levels of this rooted tree are

$$\begin{aligned} L_0 &= \{v_0\}, \\ L_1 &= \{v_1, v_4\}, \\ L_2 &= \{v_2, v_5, v_6\}, \\ L_3 &= \{v_3, v_7, v_8\}. \end{aligned}$$

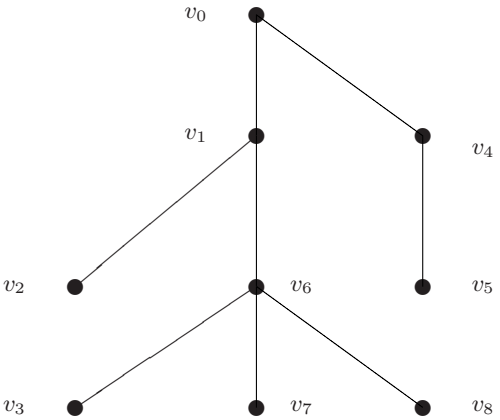



**Fig. 3.13.** Rooted tree.

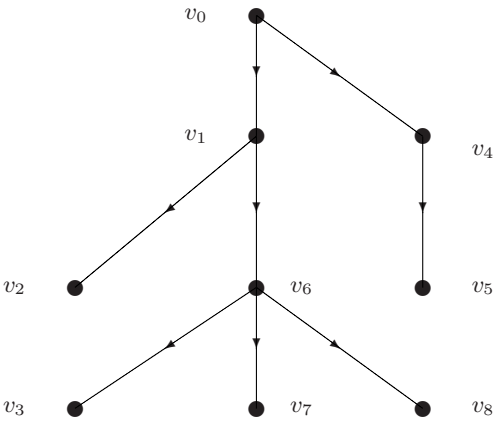



**Fig. 3.14.** Directed graph of a rooted tree.

Often we associate to each vertex of a rooted tree a sequence of its immediate descendants. The formal concept that corresponds to this idea is the notion of an *ordered rooted tree* defined as a triple $(\mathcal{T}, v_0, r)$, where $\mathcal{T} = (V, E)$ and $v_0$ have the same meaning as above, and $r : V \longrightarrow \mathbf{Seq}(V)$ is a function defined on the set of vertices of $\mathcal{T}$ such that $r(v)$ is a sequence, without repetition, of the descendants of the node $v$. If $v$ is a leaf, then $r(v) = \lambda$.

*Example 3.42.* In Figure 3.15, we present an ordered rooted tree that is created starting from the rooted tree from Figure 3.14.

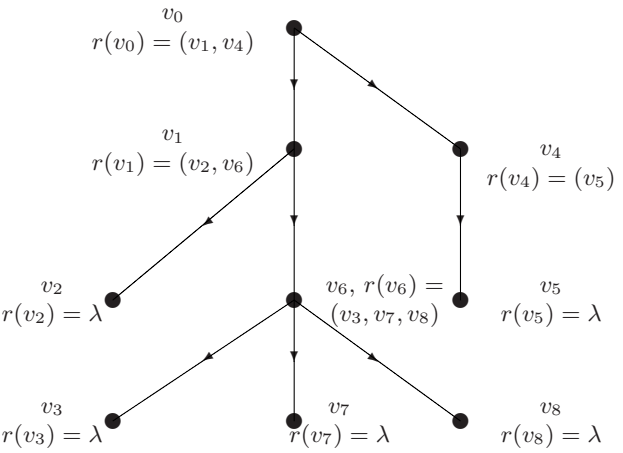



**Fig. 3.15.** Ordered rooted tree.

In general, we omit the explicit specification of the sequences of descendants for an ordered rooted tree and assume that each such sequence $r(v)$ consists of the direct descendants of $v$ read from the graph from left to right.

**Definition 3.43.** *A* binary tree *is a rooted tree* $(\mathcal{T}, v_0)$ *such that each node has at most two descendants.*

The rooted tree is a subgraph of $(\mathcal{T}, v_0)$ that consists of all descendants of the left son of $v_0$ is the *left subtree* of the binary tree. Similarly, the set of descendants of the right son of $v_0$ forms the *right subtree* of $\mathcal{T}$.

In a binary tree, a level $L_i$ may contain up to $2^i$ nodes; when this happens, we say that the level is *complete*. Thus, in a binary tree of height $h$, we may have at most $2^{h+1}$ nodes.

The notion of an *ordered binary tree* corresponds to binary trees in which we specify the order of the descendants of each node. If $v, x, y$ are three nodes of an ordered binary tree and $r(v) = (x, y)$, then we say that $x$ is the *left son of* $v$ and $y$ is the *right son of* $v$.

An *almost complete binary tree* is a binary tree such that all levels, with the possible exception of the last, are complete. The last level of an almost complete binary tree is always filled from left to right.

Note that the ratio of the number of nodes of a right subtree and the number of nodes of the left subtree of an almost complete binary tree is at most 2. Thus, the size of these subtrees is not larger than $\frac{2}{3}$ the size of the almost complete binary tree.

*Example 3.44.* The binary tree shown in Figure 3.16 is an almost complete binary tree.

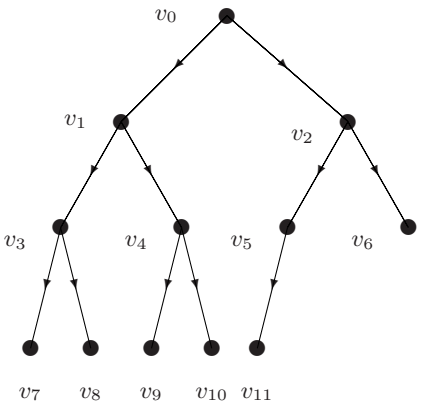



**Fig. 3.16.** An almost complete binary tree.

Almost complete binary trees are useful for defining a data structure known as a *heap*.

**Definition 3.45.** *A* heap *is a 4-tuple* $(\mathcal{T}, v_0, r, \ell)$ *such that* $(\mathcal{T}, v_0, r)$ *is an almost complete binary tree and* $\ell : V \longrightarrow \mathbb{R}$ *is a function defined on the set of vertices of* $\mathcal{T}$ *such that if* $r(v) = (v', v'')$*, then* $\ell(v) \leq \min\{\ell(v'), \ell(v'')\}$.

*Example 3.46.* Starting from the almost complete binary tree shown in Figure 3.15, we show a heap in Figure 3.17. The values of $\ell$ are placed near the names of the vertices of $(\mathcal{T}, v_0, r)$.

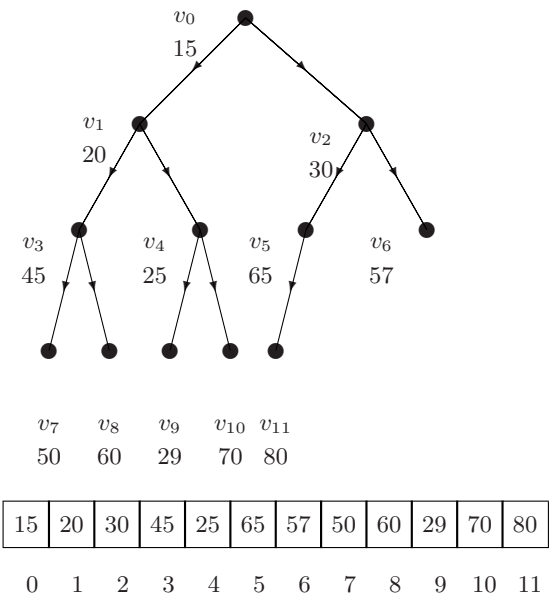



**Fig. 3.17.** A heap and its array.

A frequently used implementation of heaps makes use of arrays. The value of $\ell(v_0)$ is stored in the first element of an array $H[0]$. If the value of $\ell(v_i)$ is placed in $H[i-1]$, then the values of $\ell(v')$ and $\ell(v'')$, the left and right sons of $v$, are placed in $H[2i+1]$ and $H[2i+2]$, respectively. Thus, for a node represented in the position $j \geq 1$, its parent is placed in the position

$$\mathsf{parent}(j) = \lfloor j/2 \rfloor - 1.$$

Note that $\mathsf{parent}(j)$ is defined only if $j \geq 1$. Also, the numbers $2i+1$ and $2i+2$ are denoted by $\mathsf{leftchild}(i)$ and $\mathsf{rightchild}(i)$, respectively.

For further applications, we need to examine in some detail how a heap is put together starting from an array $A$ of size $A.\mathsf{size}$. Observe that the leafs of the heap are stored in the last $k$ nodes of the array, where $k = A.\mathsf{size} - 2^{\lfloor \log_2 A.\mathsf{size} \rfloor}$, and that the number of nonleaves is $\lfloor A.\mathsf{size}/2 \rfloor$.

The procedure called $\mathsf{heapify}(A, i)$ starts from an array $A$ and rearranges the elements of the portion of the array located between the positions $i$ and $A.\mathsf{size} - 1$ of the array such that the resulting subarray is a heap. When this procedure is called with the parameter $i$, we make the assumption that the

trees that correspond to the positions $\mathsf{leftchild}(i)$ and $\mathsf{rightchild}(i)$ are already heaps.

The procedure $\mathsf{heapify}(A, i)$ is

```
heapify(A, i){
    left = leftchild(i);
    right = rightchild(i);
    if (left < A.size and A[left] < A[i])
        then smallest = left;
    else
        smallest = right;
    if (right < A.size and A[right] < A[smallest])
        then smallest = right;
    if (smallest! = i) {
        swap(A[i],A[smallest]);
        heapify(A,smallest);
    }
}
```

The variable smallest is used to store the smallest of the elements contained by $A[i]$, $A[\mathsf{leftchild}(i)]$, and $A[\mathsf{rightchild}(i)]$. If the smallest value is contained by $A[i]$, then the algorithm halts. Otherwise, the array does not satisfy the heap condition, $A[i]$ is swapped with the smallest of its children, and the algorithm is applied to the subtree that now has $A[i]$ at its root.

As we observed, the size of a subtree cannot exceed $\frac{2n}{3}$, where $n = A.\mathsf{size}$. Thus, the time $T(n)$ required by the heapify for an array of size $n$ can be expressed as

$$T(n) = T\left(\frac{2n}{3}\right) + a,$$

where $a$ is a constant that takes into account the cost of the other operations. Applying this recurrence repeatedly, we have

$$T(n) = T\left(\left(\frac{2}{3}\right)^h n\right) + ha.$$

Let $c, d \in \mathbb{R}_{>0}$ be two fixed numbers such that $c/d > 3/2$. Choosing $h$ such that $c \leq \left(\frac{2}{3}\right)^h n \leq d$, it follows that $h = O(\log n)$, so $T(n) = O(\log n)$.

Observe that the part of the array located between positions $A.\mathsf{size} - k$ and $A.\mathsf{size} - 1$ already satisfies the heap condition since it contains only values associated to leaves (which have no children). Therefore, the algorithm that constructs the heap begins with position $\lfloor A.\mathsf{size}/2 \rfloor - 1$ and halts at position 0:

```
build_heap(A){
for i = ⌊A.size/2⌋ − 1 downto 0 do
        heapify(A, i);
}
```

The cost of this algorithm is $O(n \log n)$ since there are $O(n)$ calls to heapify and each call costs $O(\log n)$.

When using a heap, we need to define at least two operations: $\mathsf{insert}(A, x)$, which returns a heap in which the number $x$ is placed into the proper position, and $\mathsf{getsmallest}(A)$, which returns the smallest element from the heap and a new heap that contains the remaining elements. The pseudocode for the first procedure is

```
insert(A, x){
    A.size++;
    s = A.size − 1;
    A[s] = x;
    while ((s ≥ 1) and (A[s] < A[parent(s)]))
        swap(A[s], A[parent(s)]);
        s = parent(s);
    endwhile;
}
```

The procedure begins with placing the new element in the last position of the array. Then, the element percolates in the tree to its proper position through swaps with its parent. The number of swaps is proportional to the height of the tree; therefore, the time requirement of this procedure is $O(\log n)$.

The procedure $\mathsf{getsmallest}(A)$ extracts the element located at the top of the array, then places the element located at the bottom of the heap, $A[\mathsf{size} - 1]$, at the top of the heap, and then calls heapify$(A, 0)$. This procedure is

```
getsmallest(A){
z = A[0];
A[0] = A[size − 1];
A.size − −;
i = 0;
heapify(A, 0);
return z;
}
```

The cost of this procedure is again $O(\log n)$.

Given a graph $\mathcal{G} = (V, E)$ and two unconnected vertices $u, v \in V$ let $\mathcal{G} + (u, v)$ be the graph $(V, E \cup \{(u, v)\})$. If $(r, s)$ is an edge in $\mathcal{G}$, we denote by $\mathcal{G} - (r, s)$ the graph $(V, E - \{(r, s)\})$.

**Definition 3.47.** *A* weighted graph *is a pair* $(\mathcal{G}, w)$*, where* $\mathcal{G} = (V, E)$ *is a graph and* $w : E \longrightarrow \mathbb{R}$ *is a* weight function*. If* $e \in E$*, we refer to* $w(e)$ *as the* weight of the edge $e$*.*

*The* weight of a set of edges $F$ *is the number* $w(F)$ *defined by*

$$w(F) = \sum \{w(e) | e \in F\}.$$

*A* minimal spanning tree *for* $\mathcal{G} = (V, E)$ *is a spanning graph for* $\mathcal{G}$ *that is a tree* $\mathcal{T} = (V, F)$ *such that* $w(F)$ *is minimal.*

We present an effective construction of a minimal spanning tree for a weighted graph $(\mathcal{G}, w)$, where $\mathcal{G} = (V, E)$ is a finite, connected graph known as *Kruskal's algorithm*. Suppose that $e_1, e_2, \ldots, e_m$ is the list of all edges of $\mathcal{G}$ listed in increasing order of their weights; that is, $w(e_1) \leq w(e_2) \leq \ldots \leq w(e_m)$. We use the following algorithm.

**Algorithm 3.48 (Kruskal's Algorithm)**
**Input:** a weighted graph $(\mathcal{G}, w)$.
**Output:** a set of edges $F$ that defines a minimal spanning tree.
**Method:**
initialize $F = e_1$;
**repeat**
select the first edge $e$ in the list such that $e \notin F$
and the subgraph $(V, F \cup \{e\})$ is acyclic;
$F := F \cup \{e\}$;
**until** no edges exist that satisfy these conditions;
**output** $\mathcal{T} = (V, F)$

When the algorithm is completed, we built a maximal acyclic subgraph $\mathcal{T} = (V, F)$, that is, a spanning tree.

We claim that $\mathcal{T} = (V, F)$ is a minimal spanning tree. Indeed, let $\mathcal{T}' = (V, F')$ be a minimal spanning tree such that $|F \cap F'|$ is maximal. Suppose that $F' \neq F$, and let $e = (x, y)$ be the first edge of $F$ in the list of edges that does not belong to $F'$.

The tree $\mathcal{T}'$ contains a unique path $\mathbf{p}$ that joins $x$ to $y$. Note that this path cannot be included in $\mathcal{T}$ since otherwise $\mathcal{T}$ would contain a cycle formed by $\mathbf{p}$ and $(x, y)$. Therefore, there exists an edge $e'$ on the path $\mathbf{p}$ that does not belong to $\mathcal{T}$.

Note that the weight of edge $e$ cannot be larger than the weight of $e'$ because $e$ was chosen for $\mathcal{T}$ by the algorithm and $e'$ is not an edge of $\mathcal{T}$, which shows that $e$ precedes $e'$ in the previous list of edges. The set $F_1 = F' - \{e'\} \cup \{e\}$ defines a spanning tree $\mathcal{T}_1$, and since $w(F_1) = w(F') - w(e') + w(e) \leq w(F')$, it follows that the tree $\mathcal{T}' = (V, F')$ is a minimal spanning tree of $\mathcal{G}$. Since $|F_1 \cap F| > |F' \cap F|$, this leads to a contradiction. Thus, $F' = F$ and $\mathcal{T}$ is indeed a minimal spanning tree.

*Example 3.49.* Consider the weighted graph given in Figure 3.18, whose edges are marked by the weights. The list of edges in nondecreasing order of their weights is

$$(v_1, v_2), (v_1, v_4), (v_6, v_7), (v_5, v_6),$$
$$(v_3, v_4), (v_3, v_5), (v_5, v_8), (v_7, v_8),$$
$$(v_2, v_3), (v_1, v_8), (v_2, v_6), (v_4, v_6).$$

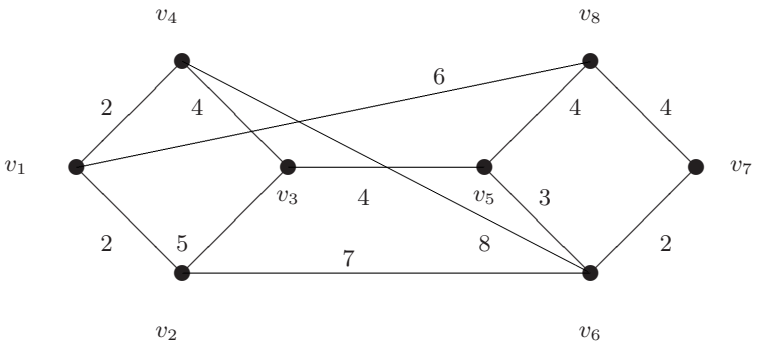



**Fig. 3.18.** Weighted graph $(\mathcal{G}, w)$.

The minimal spanning tree for this weighted graph is shown in thick lines in Figure 3.19. The sequence of edges added to the set of edges of the minimal spanning tree is

$$(v_1, v_2), (v_1, v_4), (v_6, v_7), (v_5, v_6), (v_3, v_4), (v_3, v_5), (v_5, v_8).$$

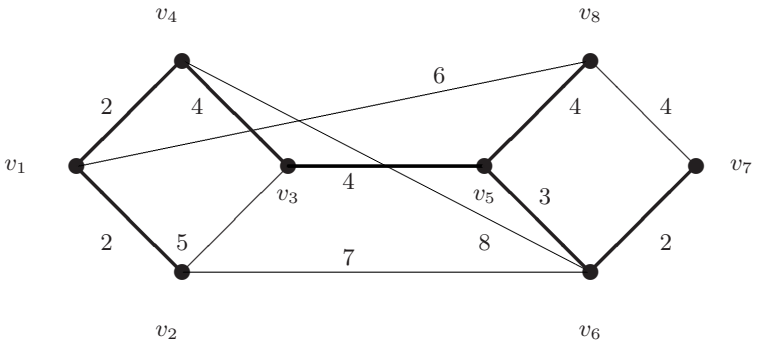



**Fig. 3.19.** Minimal spanning tree for the weighted graph from Figure 3.18.

An alternative algorithm known as *Prim's algorithm* is given next. In this modality of constructing the minimal spanning tree of a finite, connected graph $\mathcal{G} = (V, E)$, we construct a sequence of pairs of sets of vertices and edges that begins with a pair $(V_1, E_1) = (\{v\}, \emptyset)$, where $v$ is an arbitrary vertex.

Suppose that we constructed the pairs $(V_1, E_1), \ldots, (V_k, E_k)$. Define the set of edges $H_k = \{(v, w) \in E | v \in V_k, w \notin V_k\}$. If $(v_k, t_k)$ is an edge in $H_k$

of minimal weight, then $V_{k+1} = V_k \cup \{v_k\}$ and $E_{k+1} = E_k \cup \{(v_k, t_k)\}$. The algorithm halts when $H_k = \emptyset$.

Consider the increasing sequences $V_1 \subseteq V_2 \subseteq \cdots$ and $E_1 \subseteq E_2 \subseteq \cdots$. An easy induction argument on $k$ shows that the subgraphs $(V_k, E_k)$ are acyclic. The sequence halts with the pair $(V_n, E_n)$, where $H_n = \emptyset$, so $V_n = V$. Thus, $(V_n, E_n)$ is indeed a spanning tree.

To prove that $(V_n, E_n) = (V, E_n)$ is a minimal spanning tree, we will show that for every subgraph $(V_k, E_k)$, $E_k$ is a subset of the set of edges of a minimal spanning tree $\mathcal{T} = (V, E)$.

The argument is by induction on $k$. The basis case, $k = 1$, is immediate since $E_1 = \emptyset$.

Suppose that $E_k$ is a subset of the set of edges of a minimal spanning tree $\mathcal{T} = (V, E)$ and $E_{k+1} = E_k \cup \{(v_k, t_k)\}$.

Since $\mathcal{T}$ is a connected graph, there is a path in this graph that connects $v_k$ to $t_k$. Let $(r, s)$ be the first edge in this path that has one endpoint in $V_k$. By the definition of $(v_k, t_k)$, we have $w(v_k, t_k) \leq w(r, s)$. Thus, if we replace $(r, s)$ by $(v_k, t_k)$ in $\mathcal{T}$, we obtain a minimal spanning tree whose set of edges includes $E_{k+1}$.

*Example 3.50.* We apply Prim's algorithm to the weighted graph introduced in Example 3.49 starting with the vertex $v_3$.

The sequences $V_1 \subseteq V_2 \subseteq \cdots$ and $E_1 \subseteq E_2 \subseteq \cdots$ are given in the following table:

| $k$ | $E_k$ | $V_k$ |
|---|---|---|
| 1 | $\{v_3\}$ | $\emptyset$ |
| 2 | $\{v_3, v_5\}$ | $\{(v_3, v_5)\}$ |
| 3 | $\{v_3, v_5, v_6\}$ | $\{(v_3, v_5), (v_5, v_6)\}$ |
| 4 | $\{v_3, v_5, v_6, v_7\}$ | $\{(v_3, v_5), (v_5, v_6), (v_6, v_7)\}$ |
| 5 | $\{v_3, v_5, v_6, v_7, v_4\}$ | $\{(v_3, v_5), (v_5, v_6), (v_6, v_7), (v_3, v_4)\}$ |
| 6 | $\{v_3, v_5, v_6, v_7, v_4, v_1\}$ | $\{(v_3, v_5), (v_5, v_6), (v_6, v_7), (v_3, v_4), (v_4, v_1)\}$ |
| 7 | $\{v_3, v_5, v_6, v_7, v_4, v_1, v_2\}$ | $\{(v_3, v_5), (v_5, v_6), (v_6, v_7), (v_3, v_4), (v_4, v_1), (v_1, v_2)\}$ |
| 8 | $\{v_3, v_5, v_6, v_7, v_4, v_1, v_2\}$ | $\{(v_3, v_5), (v_5, v_6), (v_6, v_7), (v_3, v_4), (v_4, v_1), (v_1, v_2), (v_5, v_8)\}$ |

Next, we introduce three notions related to two-block partitions of the set of vertices of a weighted graph. These concepts will be useful in presenting an application of minimal spanning trees to clustering initiated by C. T. Zahn in [147].

**Definition 3.51.** *Let $(\mathcal{G}, w)$ be a weighted graph, where $\mathcal{G} = (V, E)$, and let $\pi = \{V_1, V_2\}$ be a two-block partition of the set of vertices $V$ of $\mathcal{G}$.*

*The* separation *of $\pi$ is*

$$sep(\pi) = \min\{w(v_1, v_2) \mid v_1 \in V_1 \text{ and } v_2 \in V_2\}.$$

*The cut set of $\pi$ is the set $CS(\pi) = \{(x, y) \in E \mid x \in V_1 \text{ and } y \in V_2\}$.*

*Finally, the* set of links *of $\pi$ is the set of edges*

$$LK(\pi) = \{(x, y) \in CS(\pi) \mid w(x, y) = sep(\pi)\}.$$

**Theorem 3.52.** *For every partition $\pi = \{V_1, V_2\}$ of a weighted graph $(\mathcal{G}, w)$ and minimal spanning tree $\mathcal{T}$, there exists an edge that belongs to $\mathcal{T}$ and to $LK(\pi)$.*

*Proof.* Suppose that $\mathcal{T}$ is a minimal spanning graph that contains no edge of $LK(\pi)$. If an edge $(v_1, v_2) \in LK(\{V_1, V_2\})$ is added to $\mathcal{T}$, the resulting graph $\mathcal{G}'$ contains a unique cycle. The part of this cycle contained in $\mathcal{T}$ must contain at least one other edge $(s,t) \in CS(\{V_1, V_2\})$ because $v_1 \in V_1$ and $v_2 \in V_2$. The edge $(s,t)$ does not belong to $LK(\{V_1, V_2\})$ by the supposition we made concerning $\mathcal{T}$. Consequently, $w(s,t) > w(v_1, v_2)$, which means that the spanning tree $\mathcal{T}_1$ obtained from $\mathcal{T}$ by removing $(s,t)$ and adding $(v_1, v_2)$ will have a smaller weight than $\mathcal{T}$. This would contradict the minimality of $\mathcal{T}$. □

**Theorem 3.53.** *If $(x,y)$ is an edge of a tree $\mathcal{T} = (V, E)$, then there exists a partition $\pi = \{V_1, V_2\}$ of $V$ such that $CS(\{V_1, V_2\}) = \{(x,y)\}$.*

*Proof.* Since $\mathcal{T}$ is a minimally connected graph, removing an edge $(x,y)$ results in a graph that contains two disjoint connected components $V_1$ and $V_2$ such that $x \in V_1$ and $y \in V_2$. Then, it is clear that $\{(x,y)\} = CS(\{V_1, V_2\})$. □

**Theorem 3.54.** *Let $(\mathcal{G}, w)$ be a weighted graph and let $\mathcal{T}$ be a minimal spanning link of $\mathcal{G}$. All minimal spanning tree edges are links of some partition of $(\mathcal{G}, w)$.*

*Proof.* Let $\mathcal{G} = (V, E)$ and let $(x,y)$ be an edge in $\mathcal{T}$. If $(V_1, V_2)$ is the partition of $V$ that corresponds to the edge $(x,y)$ that exists by Theorem 3.53, then, by Theorem 3.52, $\mathcal{T}$ must contain an edge from $CS(\{V_1, V_2\})$. Since $\mathcal{T}$ contains only one such edge, it follows that this edge must belong to $LK(\{V_1, V_2\})$. □

**Corollary 3.55.** *Let $(\mathcal{G}, w)$ be a weighted graph, where $\mathcal{G} = (V, E)$. If $w : E \longrightarrow \mathbb{R}$ is an injective mapping (that is, if all weights of the edges are distinct), then the minimal spanning tree is unique. Furthermore, this minimal spanning tree has the form $\mathcal{T} = (V, L(\mathcal{G}))$, where $L(\mathcal{G})$ is the set of all links of $G$.*

*Proof.* Let $\mathcal{T} = (V, E')$ be a minimal spanning tree of $(\mathcal{G}, w)$. The injectivity of $w$ implies that, for any partition $\pi$ of $V$, the set $LK(\pi)$ consists of a unique edge that belongs to each minimal spanning tree. Thus, $L(\mathcal{G}) \subseteq E'$. The reverse inclusion follows immediately from Theorem 3.54, so $\mathcal{G}$ has a unique spanning tree $\mathcal{T} = (V, L(\mathcal{G}))$. □

**Theorem 3.56.** *Let $(\mathcal{G}, w)$ be a weighted graph, where $\mathcal{G} = (V, E)$. If $U$ is a nonempty subset of $V$ such that $sep(\{U_1, U_2\}) < sep(U, V - U)$ for every two-block partition $\pi = \{U_1, U_2\} \in PART(U)$, then, for every minimal spanning tree $\mathcal{T}$ of $(\mathcal{G}, w)$, the subgraph $\mathcal{T}_U$ is a subtree of $\mathcal{T}$.*

*Proof.* Let $\pi = \{U_1, U_2\}$ be a two block partition of $U$. To prove the statement, it suffices to show that every minimal spanning tree $\mathcal{T}$ of $(\mathcal{G}, w)$ contains an edge in $CS(\pi)$. This, in turn, will imply that the subgraph $\mathcal{T}_U$ of $\mathcal{T}$ determined by $U$ has only one connected component, which means that $\mathcal{T}_U$ is a subtree of $\mathcal{T}$.

To prove that $\mathcal{T}$ contains an edge from $CS(\pi)$, it will suffice to show that $sep(U_1, U_2) < sep(U_1, V - U)$ because this implies $LK(U_1, V - U_1) \subseteq CS(U_1, U_2)$. Indeed, if this is the case, then the shortest link between a vertex in $U_1$ and one outside of $U_1$ must be an edge that joins a vertex from $U_1$ to a vertex in $U_2$.

Observe that

$$sep(U, V - U) = sep(U_1 \cup U_2, V - U) = \min\{sep(U_1, V - U), sep(U_2, V - U)\},$$

and therefore $sep(U_1, V - U) \geq sep(U, V - U)$.

By the hypothesis of the theorem, $sep(U, V - U) > sep(U_1, U_2)$, and therefore

$$sep(U_1, U_2) < sep(U, V - U) \leq sep(U_1, V - U),$$

which leads to the desired conclusion. □

Search enumeration trees were introduced by R. Rymon in [116] in order to provide a unified search-based framework for several problems in artificial intelligence; they are also useful for data mining algorithms.

Let $S$ be a set and let $d : S \longrightarrow \mathbb{N}$ be an injective function. The number $d(x)$ is the *index* of $x \in S$. If $P \subseteq S$, the *view of* $P$ is the subset

$$view(d, P) = \left\{ s \in S \mid d(s) > \max_{p \in P} d(p) \right\}.$$

**Definition 3.57.** *Let $\mathcal{C}$ be a hereditary collection of subsets of a set $S$. The graph $\mathcal{G} = (\mathcal{C}, E)$* is a Rymon tree for $\mathcal{C}$ and the indexing function $d$ *if*

(i) *the root of $\mathcal{G}$ is the empty set, and*

(ii) *the children of a node $P$ are the sets of the form $P \cup \{s\}$, where $s \in view(d, P)$.*

*If $S = \{s_1, \ldots, s_n\}$ and $d(s_i) = i$ for $1 \leq i \leq n$, we will omit the indexing function from the definition of the Rymon tree for $\mathcal{P}(S)$.*

*Example 3.58.* Let $S = \{i_1, i_2, i_3, i_4\}$ and let $\mathcal{C}$ be $\mathcal{P}(S)$, which is clearly a hereditary collection of sets. Define the injective mapping $d$ by $d(i_k) = k$ for $1 \leq k \leq 4$. The Rymon tree for $\mathcal{C}$ and $d$ is shown in Figure 3.20.

A key property of a Rymon tree is stated next.

**Theorem 3.59.** *Let $\mathcal{G}$ be a Rymon tree for a hereditary collection $\mathcal{C}$ of subsets of a set $S$ and an indexing function $d$. Every set $P$ of $\mathcal{C}$ occurs exactly once in the tree.*

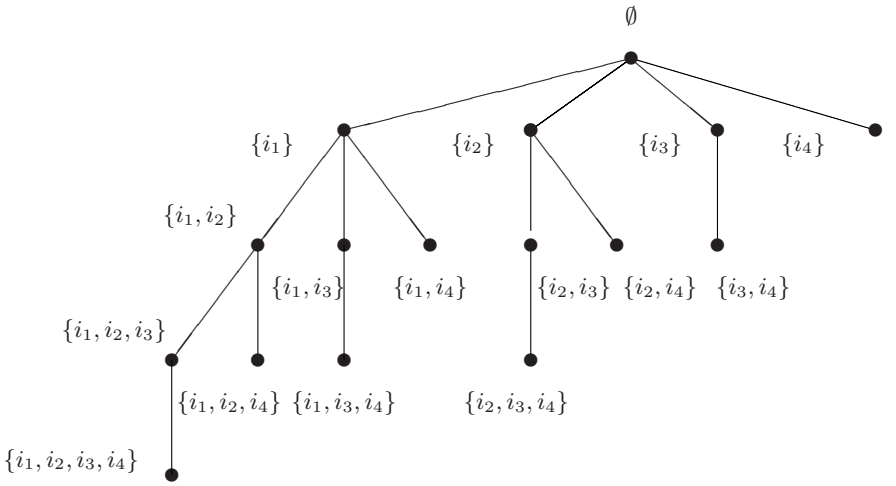



**Fig. 3.20.** Rymon tree for $\mathcal{P}(\{i_1, i_2, i_3, i_4\})$.

*Proof.* The argument is by induction on $p = |P|$. If $p = 0$, then $P$ is the root of the tree and the theorem obviously holds.

Suppose that the theorem holds for sets having fewer than $p$ elements, and let $P \in \mathcal{C}$ be such that $|P| = p$. Since $\mathcal{C}$ is hereditary, every set of the form $P - \{x\}$ with $x \in P$ belongs to $\mathcal{C}$ and, by the inductive hypothesis, occurs exactly once in the tree.

Let $z$ be the element of $P$ that has the largest value of the index function $d$. Then $\mathit{view}(P - \{z\})$ contains $z$ and $P$ is a child of the vertex $P - \{z\}$. Since the parent of $P$ is unique, it follows that $P$ occurs exactly once in the tree. □

If a set $U$ is located at the left of a set $V$ in the tree $\mathcal{G}_I$, we shall write $U \sqsubset V$. Thus, we have

$$\begin{aligned}
&\emptyset \sqsubset \{i_1\} \sqsubset \{i_1, i_2\} \sqsubset \{i_1, i_2, i_3, i_4\} \\
&\sqsubset \{i_1, i_2, i_4\} \sqsubset \{i_1, i_3\} \sqsubset \{i_1, i_3, i_4\} \\
&\sqsubset \{i_1, i_4\} \sqsubset \{i_2\} \sqsubset \{i_2, i_3\} \\
&\sqsubset \{i_2, i_3, i_4\} \sqsubset \{i_2, i_4\} \sqsubset \{i_3\} \\
&\sqsubset \{i_3, i_4\} \sqsubset \{i_4\}.
\end{aligned}$$

Note that in the Rymon tree of a collection of the form $\mathcal{P}(S)$, the collection of sets of $\mathcal{S}_r$ that consists of sets located at distance $r$ from the root denotes all $\binom{n}{r}$ subsets of size $r$ of $S$.

**Definition 3.60.** *A* numbering of a graph $\mathcal{G} = (V, E)$ *is a bijection* $\nu : V \longrightarrow 1, \ldots, |V|$. *The pair* $(\mathcal{G}, \nu)$ *is referred to as a* numbered graph.

**Theorem 3.61.** *Let* $\nu : V \longrightarrow \{1, \ldots, n\}$ *be a bijection on the set* $V$, *where* $|V| = n$. *There are* $n^{n-2}$ *numbered trees* $(\mathcal{T}, \nu)$ *having* $V$ *as the set of vertices.*

*Proof.* The best-known argument for this theorem is based on a bijection between the set of numbered trees having $n$ vertices and the set of sequences of length $n-2$ defined on the set $\{1, \ldots, n\}$ and has been formulated in [109].

Let $(\mathcal{T}, \nu)$ be a numbered tree having $n$ vertices. Define a sequence of trees $(\mathcal{T}_1, \ldots, \mathcal{T}_{n-1})$ and a *Prüfer sequence* $(\ell_1, \ldots, \ell_{n-2}) \in \mathbf{Seq}_n(\mathbb{N})$ as follows. The initial tree $\mathcal{T}_1$ equals $\mathcal{T}$. The tree $\mathcal{T}_i$ will have $n-i+1$ vertices for $1 \leq i \leq n-1$.

The $\mathcal{T}_{i+1}$ is obtained from $\mathcal{T}_i$ by seeking the leaf $x$ of $\mathcal{T}_i$ such that $\nu(x)$ is *minimal* and deleting the unique edge of the form $(x, y)$. The number $\nu(y)$ is added to the Prüfer sequence. Note that the label $\ell$ of a vertex $u$ will occur exactly $\mathsf{d}(u) - 1$ times in the Prüfer sequence, once for every vertex adjacent to $u$ that is removed in the process of building the sequence of trees.

Let $L(\mathcal{T}, \nu)$ be the Prüfer sequence of $(\mathcal{T}, \nu)$. If $NT_n$ is the set of numbered trees on $n$ vertices, then the mapping $L : NT_n \longrightarrow \mathbf{Seq}_{n-2}(\{1, \ldots, n\})$ is a bijection.

The edges that are removed in the process of constructing the Prüfer sequences can be listed in a table:

| Starting Tree | Leaf | Vertex | Resulting Tree |
|---|---|---|---|
| $\mathcal{T}_1$ | $x_1$ | $y_1$ | $\mathcal{T}_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\mathcal{T}_{n-2}$ | $x_{n-2}$ | $y_{n-2}$ | $\mathcal{T}_{n-1}$ |
| $\mathcal{T}_{n-1}$ | $x_{n-1}$ | $y_{n-1}$ | $-$ |

Note that the edges of $\mathcal{T}_i$ are $(x_j, y_j)$ for $i \leq j \leq n-1$.

The next to the last tree in the sequence, $\mathcal{T}_{n-2}$, has two edges and therefore three vertices. The last tree in the sequence $\mathcal{T}_{n-1}$ consists of a unique edge $(x_{n-1}, y_{n-1})$. Since a tree with at least two vertices has at least two leaves, the node whose label is $n$ will never be the leaf with the minimal label. Therefore, $\nu(y_{n-1}) = n$ and $n$ is always the last number of $L(\mathcal{T}, \nu)$.

Also, observe that the leaves of the tree $\mathcal{T}_i$ are those vertices that do not belong to $\{x_1, \ldots, x_{i-1}, y_i, \ldots, y_{n-1}\}$, which means that $x_i$ is the vertex that has the minimal label and is not in the set above. In particular, $x_1$ is the vertex that has the least label and is not in $L(\mathcal{T}, \nu)$. This shows that we can uniquely determine the vertices $x_i$ from $L(\mathcal{T}, \nu)$ and $x_1, \ldots, x_{i-1}$. □

*Example 3.62.* Consider the tree $\mathcal{T}$ shown in Figure 3.21.

The labels of the vertices are placed at the right of each rectangle that represents a vertex. The table that contains the sequence of edges is

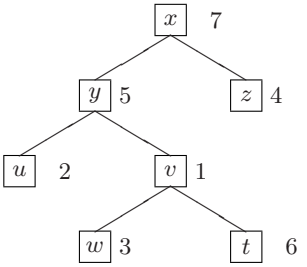



**Fig. 3.21.** Enumerated tree.

| Starting Tree | Leaf | Vertex $y$ | $\nu(y)$ | Resulting Tree |
|---|---|---|---|---|
| $\mathcal{T}_1$ | $u$ | $y$ | 5 | $\mathcal{T}_2$ |
| $\mathcal{T}_2$ | $w$ | $v$ | 1 | $\mathcal{T}_3$ |
| $\mathcal{T}_3$ | $z$ | $x$ | 7 | $\mathcal{T}_4$ |
| $\mathcal{T}_4$ | $t$ | $v$ | 1 | $\mathcal{T}_5$ |
| $\mathcal{T}_5$ | $v$ | $y$ | 5 | $\mathcal{T}_6$ |
| $\mathcal{T}_6$ | $y$ | $x$ | 7 | – |

This means that $L(\mathcal{T}, \mu) = (5, 1, 7, 1, 5) = \nu^{-1}(y, v, x, v, y)$. The vertex with the smallest label that does occur in $L(\mathcal{T}, \mu) = (5, 1, 7, 1, 5)$ is $u$ because $\ell(u) = 2$. This means that the first edge is $(u, y)$ since $\ell(y) = 5$. The succession of trees is shown in Figure 3.22. Under each tree, we show the sequence $(x_1, \ldots, x_{i-1}, y_i, \ldots, y_{n-1})$, which allows us to select the current leaf $x_i$.

*Example 3.63.* Suppose again that we have a tree having the set of nodes $V = \{x, y, z, u, v, w, t\}$ with the numbering given by

| Vertex | $x$ | $y$ | $z$ | $u$ | $v$ | $w$ | $t$ |
|---|---|---|---|---|---|---|---|
| $\nu$(vertex) | 7 | 5 | 4 | 2 | 1 | 3 | 6 |

We reconstruct the tree that has $(2, 3, 5, 2, 3)$ as its Prüfer sequence. The first leaf of this tree will be the vertex with the least value of $\nu$ that is not present in the sequence $\nu^{-1}(2, 3, 5, 2, 3) = (u, w, y, u, w)$; that is, $v$. This means that we start with the following table.

| Starting Tree | Leaf | Vertex $y$ | $\nu(y)$ | Resulting Tree |
|---|---|---|---|---|
| $\mathcal{T}_1$ | $x$ | $u$ | 2 | $\mathcal{T}_2$ |
| $\mathcal{T}_2$ | | $w$ | 3 | $\mathcal{T}_3$ |
| $\mathcal{T}_3$ | | $y$ | 5 | $\mathcal{T}_4$ |
| $\mathcal{T}_4$ | | $u$ | 2 | $\mathcal{T}_5$ |
| $\mathcal{T}_5$ | | $w$ | 3 | $\mathcal{T}_6$ |
| $\mathcal{T}_6$ | | $x$ | 7 | – |

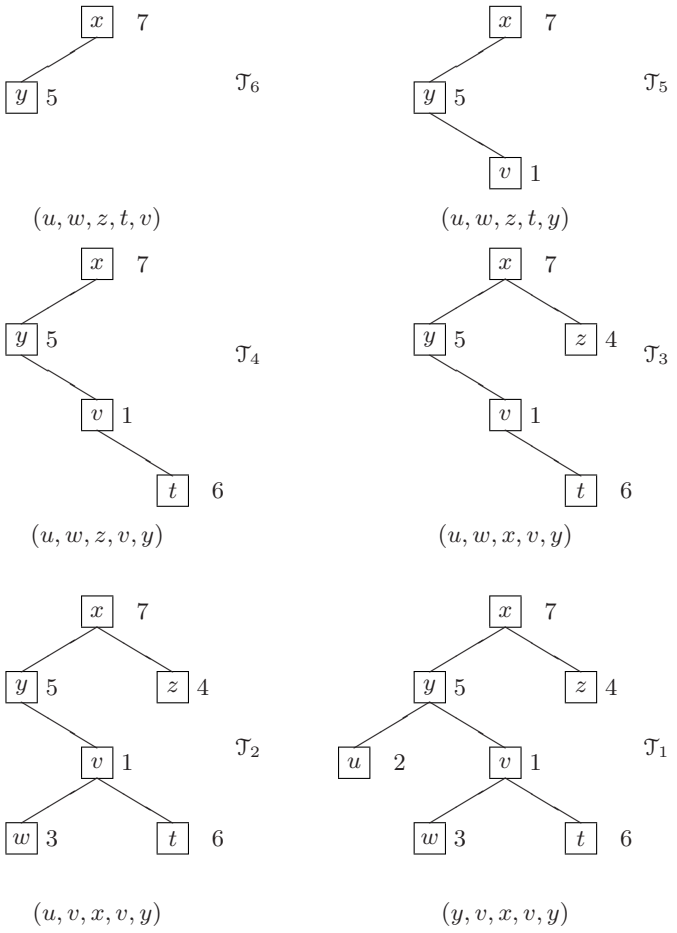



**Fig. 3.22.** Numbered trees with sequences $(x_1, \ldots, x_{i-1}, y_i, \ldots, y_{n-1})$.

For each step in filling in this table, we construct the sequence

$$(x_1, \ldots, x_{i-1}, y_i, \ldots, y_{n-1})$$

and choose $x_i$ as the vertex having minimal numbering that is not in the sequence. The final table is

| Starting Tree | Leaf | Vertex $y$ | $\nu(y)$ | Resulting Tree |
|---|---|---|---|---|
| $\mathcal{T}_1$ | $x$ | $u$ | 2 | $\mathcal{T}_2$ |
| $\mathcal{T}_2$ | $z$ | $w$ | 3 | $\mathcal{T}_3$ |
| $\mathcal{T}_3$ | $t$ | $y$ | 5 | $\mathcal{T}_4$ |
| $\mathcal{T}_4$ | $y$ | $u$ | 2 | $\mathcal{T}_5$ |
| $\mathcal{T}_5$ | $u$ | $w$ | 3 | $\mathcal{T}_6$ |
| $\mathcal{T}_6$ | $w$ | $x$ | 7 | – |

and gives the tree shown in Figure 3.23.

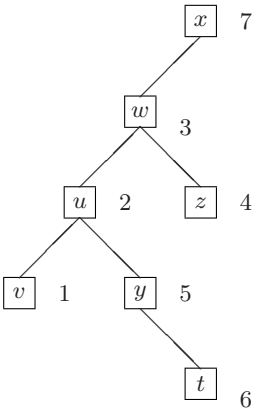



**Fig. 3.23.** Tree reconstructed from its Prüfer sequence.

## 3.4 Flows in Digraphs

**Definition 3.64.** *A* network *is a* 4*-tuple* $\mathcal{N} = (\mathcal{G}, cap, s, t)$*, where*

- $\mathcal{G} = (V, E)$ *is a finite digraph,*
- $cap : V \times V \longrightarrow \mathbb{R}_{\geq 0}$ *is a function called the* capacity function *such that* $(u, v) \notin E$ *implies* $cap(u, v) = 0$*, and*
- $s$ *and* $t$ *are two distinct vertices of* $\mathcal{G}$*, referred to as the* source *and the* sink*, respectively.*

*The number* $cap(e)$ *is the* capacity of the edge $e$*. If* $\boldsymbol{p} = (v_0, \ldots, v_n)$ *is a path in the graph* $\mathcal{G}$ *the capacity of this path is the number* $cap(\boldsymbol{p}) = \min\{cap(v_i, v_{i+1}) \mid 0 \leq i \geq n-1\}$.

*Example 3.65.* The network $\mathcal{N} = (\mathcal{G}, cap, s, t)$ is shown in Figure 3.24. If $(u, v)$ is an edge in $\mathcal{G}$, the number $cap(u, v)$ is written near the edge.

The capacity of the path $\mathbf{p} = (v_1, v_2, v_5, v_6)$ is 3 because the smallest capacity of an edge on this path is $cap(v_2, v_5) = 3$.

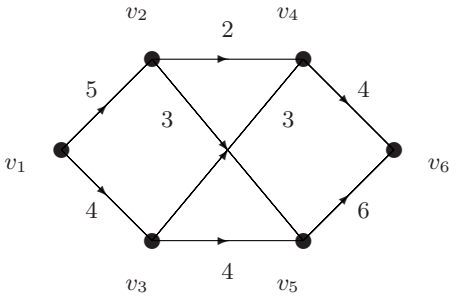



**Fig. 3.24.** 6-vertex network.

**Definition 3.66.** *A* flow *in the network* $\mathcal{N} = (\mathcal{G}, cap, s, t)$ *is a function* $f : V \times V \longrightarrow \mathbb{R}$ *that satisfies the following conditions:*

(i) *For every edge* $(u, v) \in E$ *we have* $0 \leq f(u, v) \leq cap(u, v)$.

(ii) *The function* $f$ *is skew-symmetric, that is,* $f(u, v) = -f(v, u)$ *for every pair* $(u, v) \in V \times V$.

(iii) *The equality*

$$\sum\{f(v, x) \mid v \in V\} = 0,$$

*known as* Kirkhoff's law*, holds for every vertex* $x \in V - \{s, t\}$.

*The* value of a flow $f$ *in a network* $\mathcal{N} = (\mathcal{G}, cap, s, t)$ *is the number*

$$val(f) = \sum\{f(s, v) \mid v \in V\},$$

*that is, the net flow that exits the source.*

*The set of flows of a network* $\mathcal{N}$ *is denoted by* $FL(\mathcal{N})$.

*A flow* $h$ *in* $\mathcal{N}$ *is* maximal *if* $val(f) \leq val(h)$ *for every flow* $f \in FL(\mathcal{N})$.

*If* $f(u, v) = c(u, v)$ *for an edge of* $\mathcal{G}$*, then we say that the edge* $(u, v)$ *is* saturated.

Let $\mathcal{N} = (\mathcal{G}, cap, s, t)$ be a network, where $\mathcal{G} = (V, E)$. If $f(u, v) = 0$ for every pair $(u, v) \in V \times V$, then $f$ is a flow in the network. We will refer to this flow as the *zero flow* in $\mathcal{N}$.

**Theorem 3.67.** *Let* $\mathcal{N} = (\mathcal{G}, cap, s, t)$ *be a network,* $f$ *and* $g$ *be two flows in* $FL(\mathcal{N})$*, and* $a$ *and* $b$ *be two real numbers. Define the function* $af + bg$ *by*

$$(af + bg)(u, v) = af(u, v) + bg(u, v)$$

*for every* $(u, v) \in V \times V$*. If* $0 \leq (af + bg)(u, v) \leq c(u, v)$ *for* $(u, v) \in V \times V$*, then* $af + bg$ *is a flow in the network* $\mathcal{N}$.

*Proof.* We leave this easy argument to the reader. □

Note that if $v \in V - \{s, t\}$, Kirkhoff's law can be written equivalently as

$$\sum\{f(v,x) \mid v \in V, (v,x) \in E\} + \sum\{f(v,x) \mid v \in V, (x,v) \in E\} = 0,$$

which amounts to

$$\sum\{f(v,x) \mid v \in V, (v,x) \in E\} = \sum\{f(x,v) \mid v \in V, (x,v) \in E\}$$

due to the skew symmetry of $f$.

**Theorem 3.68.** *Let $\mathcal{N} = (\mathcal{G}, cap, s, t)$ be a network and let $f$ be a flow in $\mathcal{N}$. If $U$ is a set of vertices such that $s \in U$ and $t \notin U$, then*

$$\sum\{f(u,v) \mid (u,v) \in out(U)\} - \sum\{f(u,v) \mid (u,v) \in in(U)\} = val(f).$$

*Proof.* If $f$ is a flow in $\mathcal{N}$ and $x$ is a vertex in $\mathcal{G}$, then

$$\sum\{f(x,v) \mid v \in V\} - \sum\{f(v,x) \mid v \in V\} = \begin{cases} val(f) & \text{if } x = s, \\ 0 & \text{if } x \in U - \{s\}. \end{cases}$$

Therefore,

$$\sum_{x \in U} \Big( \sum\{f(x,v) \mid v \in V\} - \sum\{f(v,x) \mid v \in V\} \Big) = val(f).$$

If an edge that occurs in the inner sums has both its endpoints in $U$, then its contribution in the first inner sum cancels with its contribution in the second inner sum. Thus, the previous equality can be written as

$$\sum\{f(u,v) \mid (u,v) \in out(U)\} - \sum\{f(u,v) \mid (u,v) \in in(U)\} = val(f).$$

□

**Corollary 3.69.** *Let $\mathcal{N} = (\mathcal{G}, cap, s, t)$ be a network and let $f$ be a flow in $\mathcal{N}$. For every vertex $x$, we have*

$$\sum\{f(v,x) \mid v \in V\} - \sum\{f(x,v) \mid v \in V\} = val(f).$$

*Proof.* Choose $U = V - \{x\}$. For the set $U$, we have

$$out(U) = (V \times \{x\}) \cap E,$$
$$in(U) = (\{x\} \times V) \cap E,$$

so it follows that

$$\sum\{f(v,x) \mid v \in v\} - \sum\{f(x,v) \mid v \in V\} = val(f).$$

□

**Definition 3.70.** *A* cut *in a network* $\mathcal{N} = (\mathcal{G}, cap, s, t)$ *is a partition* $\{C, C'\}$ *of the set of vertices* $V$ *of the digraph* $\mathcal{G}$ *such that* $s \in C$ *and* $t \in C'$.

*The* capacity of a cut $(C, C')$ *is the number*

$$cap(C, C') = \sum\{cap(u, w) \mid u \in C \text{ and } w \in C'\}.$$

*A cut with minimal capacity is a* minimal cut.

*If* $f$ *is a flow in* $\mathcal{N}$ *and* $(C, C')$ *is a cut, the* value of the flow $f$ across the cut $(C, C')$ *is the number*

$$f(C, C') = \sum\{f(u, w) \mid u \in C \text{ and } w \in C'\}.$$

*The set of cuts of a network* $\mathcal{N}$ *is denoted by* $CUTS(\mathcal{N})$.

Thus, Theorem 3.68 can be rephrased as stating that the flow across any cut equals the value of the flow. An essential observation is that since the value of a flow across a cut cannot exceed the capacity of the cut, it follows that the value of any flow is less than the capacity of any cut. As we shall see, the maximum value of a flow equals the minimal capacity of a cut.

**Definition 3.71.** *Let* $\mathcal{N} = (\mathcal{G}, cap, s, t)$ *be a network and let* $f$ *be a flow in* $\mathcal{N}$. *The* residual network of $\mathcal{N}$ relative to $f$ *is the network* $RES(\mathcal{N}, f) = (\mathcal{N}, cap', s, t)$, *where* $cap'(u, v) = cap(u, v) - f(u, v)$.

**Theorem 3.72.** *Let* $f$ *be a flow on the network* $\mathcal{N} = (\mathcal{G}, cap, s, t)$ *and let* $g$ *be a maximal flow of* $\mathcal{N}$. *The value of a maximal flow on* $RES(\mathcal{N}, f)$ *is* $val(g) - val(f)$.

*Proof.* Let $f'$ be a flow in the residual network $RES(\mathcal{N}, f)$. It is easy to see that $f + f' \in FL(\mathcal{N})$, so $val(f') \leq val(g) - val(f)$. On the other hand, $h = g - f$ is a flow on $RES(\mathcal{N}, f)$ and $val(h) = val(g) - val(f)$, so $h$ is a maximal flow having the value $val(g) - val(f)$. □

**Theorem 3.73.** *The following statements that concern a flow* $f$ *in a network* $\mathcal{N} = (\mathcal{G}, cap, s, t)$ *are equivalent:*

(i) $f$ *is a maximal flow.*
(ii) *There is no path that joins* $s$ *to* $t$ *in the residual network* $RES(\mathcal{N}, f)$ *with a positive capacity.*
(iii) $val(f) = cap(C, C')$ *for some cut* $(C, C')$ *in* $\mathcal{N}$.

*Proof.* (i) implies (ii): Let $f$ be a maximal flow in $\mathcal{N}$, and suppose that there is a path $\mathbf{p}$ in the residual network $RES(\mathcal{N}, f)$ with a positive capacity. Then, the flow $g$ defined by

$$g(u, v) = \begin{cases} f(u, v) + cap(\mathbf{p}) & \text{if } (u, v) \text{ is an edge on } \mathbf{p}, \\ f(u, v) & \text{otherwise,} \end{cases}$$

is a flow in $\mathcal{N}$ and $val(g) = val(f) + cap(\mathbf{p})$, which contradicts the maximality of the flow $f$.

(ii) implies (iii): Suppose that there is no path is the residual network $RES(\mathcal{N}, f)$ that joins the source with the sink and has a positive capacity.

Let $C$ be the set of vertices that can be reached from $s$ via a path with positive capacity (to which we add $s$) in the residual network $RES(\mathcal{N}, f)$ and let $C' = V - C$. Then the pair $(C, C')$ is a cut in $\mathcal{N}$. Observe that if $x \in C$ and $y \in C'$, then the residual capacity of the edge $(x, y)$ is 0 by the definition of $C$, which means that $f(x, y) = cap(x, y)$. Thus, $val(f) = f(C, C')$.

(iii) implies (i): Since any flow value is less than the capacity of any cut, it is immediate that $f$ is a maximal flow. □

Any path that joins the source to the target of a network $\mathcal{N}$ and has a positive capacity in that network is called an *augmenting path for the network.*

Theorem 3.73 suggests the following algorithm for constructing a maximal flow in a network.

**Algorithm 3.74 (The Ford-Fulkerson Algorithm)**
**Input:** a network $\mathcal{N} = (\mathcal{G}, cap, s, t)$.
**Output:** a maximal flow in $\mathcal{N}$.
**Method:**
initialize flow $f$ to the zero flow in $\mathcal{N}$;
**while** (there exists an augmenting path $\mathbf{p}$) **do**
augment flow $f$ along $\mathbf{p}$;
**return** $f$

*Example 3.75.* To find a maximal flow, we begin with the zero flow $f_0$ shown in Figure 3.25(a). There are several cuts in this graph having a minimal capacity equal to 9. One such cut is $\{\{v_1\}, \{v_2, v_3, v_4, v_5, v_6\}\}$; the edges that join the two sets are $(v_1, v_2)$ and $(v_1, v_3)$, which have the capacities 4 and 5, respectively.

The first augmenting path is $(v_1, v_2, v_4, v_6)$, having capacity 2. The flow $f_1$ along this path has the value 2, it saturates the edge $(v_2, v_4)$, and is shown in Figure 3.25(b). The next augmenting path is $(v_1, v_2, v_5, v_6)$, which has a capacity of 3 and is shown in Figure 3.25(c). Now the edges $(v_1, v_2)$ and $(v_2, v_5)$ become saturated. The next flow in also shown in Figure 3.25(c). In Figure 3.25(d), we show the augmenting path $(v_1, v_3, v_5, v_6)$ having capacity 3. This saturates the edge $(v_5, v_6)$. Finally, the last augmenting path of capacity 1 (shown in Figure 3.25(e)) is $(v_1, v_3, v_4, v_6)$, and the corresponding flow saturates the edge $(v_1, v_3)$. The value of the flow is 9.

**Corollary 3.76.** *Let $\mathcal{N} = (\mathcal{G}, cap, s, t)$ be a network such that $cap(e)$ is an integer for every edge of the graph $\mathcal{G}$. Then, a maximal flow ranges over the set of natural numbers.*

*Proof.* The argument for the corollary is on the number $n$ of augmenting paths used for the construction of a flow.

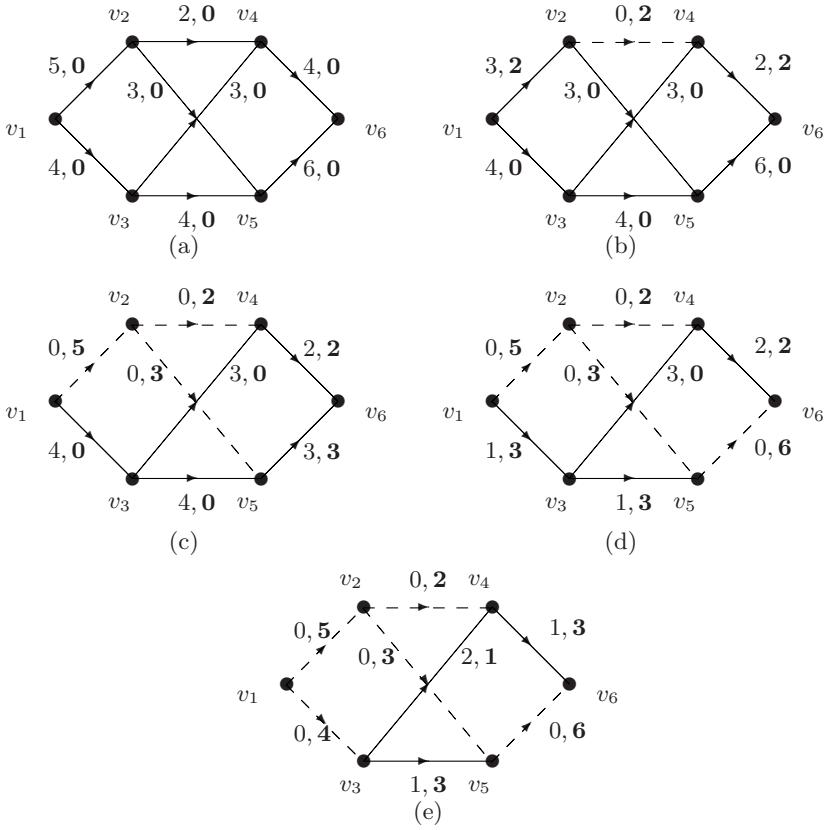



**Fig. 3.25.** Construction of a flow in a network.

The basis step, $n = 0$, is immediate since the zero flow takes as values natural numbers.

Suppose that the statement holds for the flow constructed after applying $n - 1$ augmentations. Since the value of any path is a natural number and the residual capacities are integers, the values obtained after using the $n$th augmenting path are again integers. □

Flows in networks that range over the set of integers are referred to as *integral flows*.

Network flows can be used to prove several important graph-theoretical results.

We begin with a technical result.

**Lemma 3.77.** *Let $\mathcal{N} = (\mathcal{G}, cap, x, y)$ be a network such that $cap(u, v) = 1$ for every edge $(u, v) \in E$. If $f$ is an integral flow in $FL(\mathcal{N})$ and $val(f) = m$, then there are $m$ pairwise edge-disjoint simple paths from $x$ to $y$.*

*Proof.* If $f$ is a flow in $\mathcal{N}$ that ranges over integers, then $f(u, v) \in \{0, 1\}$ for every edge $(u, v) \in E$. Also, note that the capacity of any path that joins $x$ to $y$ equals 1.

Let $E_f$ be the set of edges saturated by the flow $f$,

$$E_f = \{(u, v) \in E \mid f(u, v) = 1\}.$$

Note that no two distinct simple paths can share an edge because this would violate Kirkhoff's law. Since each path contributes a unit of flow to $f$, it follows that there exist $m$ pairwise edge-disjoint paths in $\mathcal{N}$. □

**Theorem 3.78 (Menger's Theorem).** *Let $\mathcal{G} = (V, E)$ be a directed graph and let $x$ and $y$ be two vertices. The maximum number of paths that join $x$ to $y$ whose sets of edges are pairwise disjoint equals the minimum number of edges whose removal eliminates all paths from $x$ to $y$.*

*Proof.* Define a network $\mathcal{N} = (\mathcal{G}, cap, x, y)$ such that $cap(u, v) = 1$ for every edge $(u, v) \in E$, and let $f$ be an maximal integral flow in $\mathcal{N}$ whose value is $m$. By Lemma 3.77, this number equals the number of pairwise edge-disjoint simple paths from $x$ to $y$ and is also equal to the minimum capacity of a cut in $\mathcal{N}$. Since this latter capacity equals the minimal number of edges whose removal eliminates all paths from $x$ to $y$, we obtain the desired equality. □

Menger's theorem allows us to prove the following statement involving bipartite graphs.

**Definition 3.79.** *Let $\mathcal{G} = (V, E)$ be a bipartite graph. A* matching *of $\mathcal{G}$ is a set $M$ of edges such that no two distinct edges have a common endpoint.*

*An* edge cover *of $\mathcal{G}$ is a set of vertices $U$ such that, for every edge $e \in V$, one of its endpoints is in $U$.*

**Theorem 3.80 (König's Theorem for Bipartite Graphs).** *Let $\mathcal{G} = (V, E)$ be a bipartite graph. A maximum size of a matching of $\mathcal{G}$ equals the minimum size of an edge cover of $\mathcal{G}$.*

*Proof.* Suppose that $\{V_1, V_2\}$ is the partition of the vertices of the graph such that $E \subseteq V_1 \times V_2$. Define the digraph $\mathcal{G}' = (V \cup \{s, t\}, \{(x, y) \mid (x, y) \in E\} \cup \{(s, x) \mid x \in V\} \cup \{(x, t) \mid x \in V\})$ and the network $\mathcal{N} = (\mathcal{G}', cap, s, t)$. We assume that $s, t$ are new vertices. The capacity of every edge equals 1.

A matching $M$ of the bipartite graph $\mathcal{G}$ yields a number of $|M|$ pairwise disjoint paths in $\mathcal{N}$.

An edge cover $U$ in $\mathcal{G}$ produces a cut in the network $\mathcal{N}$ using the following mechanism. Define the sets of vertices $U_1, U_2$ as $U_i = U \cap V_i$ for $i \in \{1, 2\}$. Then, $(\{s\} \cup U_1, U_2 \cup \{t\})$ is a cut in $\mathcal{N}$ and its capacity equals $|U|$. By removing the edges having an endpoint in $U$, we eliminate all paths that join $s$ to $t$. Thus, the current statement follows immediately from Menger's theorem. □

## 3.5 Hypergraphs

**Definition 3.81.** *A* hypergraph *is a pair* $\mathcal{H} = (V, \mathcal{E})$ *such that*
(i) $V$ *is a set and*
(ii) $\mathcal{E}$ *is a collection of nonempty subsets of* $V$.
*If* $\bigcup \mathcal{E} = V$, *then we say that* $\mathcal{H}$ *is a* full *hypergraph. The hypergraph is* simple *if* $X, Y \in \mathcal{E}$ *and* $X \subseteq Y$ *implies* $X = Y$. *A simple hypergraph is also known as* Sperner family of sets.

*The elements of* $V$ *are the* vertices of the hypergraph, *while the sets of* $\mathcal{E}$ *are the* hyperedges of the hypergraph.

*The* order *of the hypergraph* $\mathcal{H} = (V, \mathcal{E})$ *is* $|V|$.

Note that a graph is a simple hypergraph whose hyperedges contain two vertices.

Hyperedges are drawn as curves encircling the vertices they contain.

*Example 3.82.* The hypergraph $\mathcal{H} = (\{v_1, \ldots, v_9\}, \{U_1, \ldots, U_5\})$, where

$$\begin{aligned} U_1 &= \{v_1, v_2, v_4\}, \\ U_2 &= \{v_3\}, \\ U_3 &= \{v_4, v_7, v_8, v_9\}, \\ U_4 &= \{v_5, v_8, v_9\}, \\ U_5 &= \{v_5, v_6\}, \end{aligned}$$

is represented in Figure 3.26.

The *incidence matrix of a hypergraph* $\mathcal{H} = (\{v_1, \ldots, v_m\}, \{U_1, \ldots, U_n\})$ is an $m \times n$ matrix $M$, where

$$M_{ij} = \begin{cases} 1 & \text{if } v_i \in U_j, \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq i \leq m$ and $1 \leq j \leq n$.

*Example 3.83.* The incidence matrix of the hypergraph considered in Example 3.82 is the $9 \times 5$ matrix

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

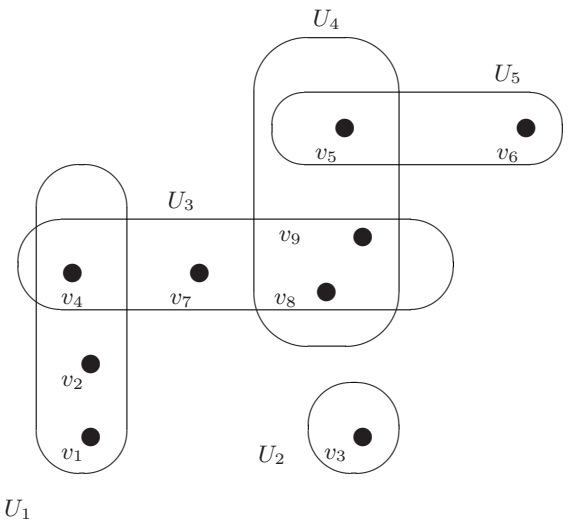



**Fig. 3.26.** Graphical representation of a hypergraph.

The definition of a hypergraph implies that every column of the incidence matrix contains at least a 1 (because hyperedges are nonempty subsets of vertices). Also, if $\mathcal{H}$ is a full hypergraph, then each row must contain at least one 1.

**Theorem 3.84.** *If $\mathcal{H} = (V, \mathcal{E})$ is a simple hypergraph of order $n$, then we have the inequality*

$$\sum_{U \in \mathcal{E}} \frac{1}{\binom{n}{|U|}} \leq 1. \tag{3.2}$$

*Moreover, we have*

$$|\mathcal{E}| \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}. \tag{3.3}$$

*Proof.* Let $V$ be a finite set such that $|S| = n$. Define the digraph $\mathcal{G}_S = (\mathcal{P}(V), E)$ having the set of edges

$$E = \{(T, U) \in \mathcal{P}(V) \times \mathcal{P}(V) \mid T \subset U \text{ and } |T| = |U| - 1\}.$$

If $U \in \mathcal{E}$, the number of paths in $\mathcal{G}$ from $\emptyset$ to $U$ is $|U|!$. Note that there are $|U|!(n - |U|)!$ paths from $\emptyset$ to $V$ that pass through $U$. If $U$ is a hyperedge of $\mathcal{H}$, then no such path contains another hyperedge of the graph because $\mathcal{H}$ is a simple hypergraph. Therefore, we have

$$n! \geq \sum_{U \in \mathcal{E}} |U|!(n - |U|)!,$$

which gives Inequality (3.2).

For the second inequality, observe that $\binom{n}{|U|} \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}$. Therefore,

$$\frac{|\mathcal{E}|}{\binom{n}{\lfloor \frac{n}{2} \rfloor}} \leq \frac{1}{\binom{n}{|U|}} \leq 1,$$

which implies Inequality (3.3). □

**Definition 3.85.** *A* hitting set *(or a* transversal*) of a hypergraph $\mathcal{H} = (V, \mathcal{E})$ is a subset $T$ of $V$ such that $T \cap E \neq \emptyset$ for every hyperedge $E$ of $\mathcal{H}$.*

*$T$ is a* minimal transversal (or a minimal hitting set) of $\mathcal{H}$ *if $T$ is a transversal and for no $U \subset T$ is $U$ a transversal (or a hitting set).*

The set of minimal transversals of a hypergraph $\mathcal{H}$ denoted by $\mathcal{M}(\mathcal{H})$ defines a simple hypergraph $MINTR(\mathcal{H}) = (V, \mathcal{M}(\mathcal{H}))$.

**Theorem 3.86.** *Let $\mathcal{H} = (V, \mathcal{E})$ and $\mathcal{H}' = (V, \mathcal{E}')$ be two simple hypergraphs on the set $V$. Then $\mathcal{H}' = MINTR(\mathcal{H})$ if and only if for every two-block partition $\pi = \{A, B\}$ of $V$ exactly one of the following situations occurs:*
(i) *There exists an $E \in \mathcal{E}$ such that $E \subseteq A$.*
(ii) *There exists an $E' \in \mathcal{E}'$ such that $E' \subseteq B$.*

*Proof.* Suppose that $\mathcal{H}' = MINTR(\mathcal{H})$, and consider a two-block partition $\pi = \{A, B\} \in PART(V)$.

If there exists $U \in \mathcal{E}$ such that $U \subseteq A$, then we have (i); otherwise, $B = V - A$ intersects each $U \in \mathcal{E}$, which means that $B$ is a transversal of $\mathcal{H}$ and therefore contains a minimum transversal $U'$ of $\mathcal{H}$. Thus, $U'$ is a hyperedge of $\mathcal{H}'$ and $U' \subseteq B$ and we have (ii).

Note that (i) and (ii) cannot occur together for if we were to have $U \in \mathcal{E}$ such that $U \subseteq A$ and $U' \in \mathcal{E}'$ such that $U' \subseteq B$, then $U' \cap U = \emptyset$.

To prove the reverse implication, suppose that the condition of the theorem holds, and let $U'$ be a hyperedge of $\mathcal{H}'$. Suppose that there exists a hyperedge $U$ of $\mathcal{H}$ such that $U' \cap U = \emptyset$, that is, $U \subseteq V - U'$. By applying the condition of the theorem to the partition $\{U', V - U'\}$, it follows that there is no $U'_1 \in \mathcal{E}'$ such that $U'_1 \subseteq U'$, which is absurd (because the role of $U'_1$ can be played by $U'$). Thus, $U'$ is a transversal of $\mathcal{H}$.

We claim that $U'$ is a minimal transversal. Suppose that the transversal $U'$ is not minimal; then, it strictly includes a minimal transversal $G$ of $\mathcal{H}$. Consider the partition $\{V - G, G\}$. Suppose that there exists $U \in \mathcal{E}$ such that $U \subseteq V - G$; this leads to a contradiction because $U$ and $G$ should have a nonempty intersection since $G$ is a transversal of $\mathcal{H}$. The alternative is the existence of a hyperedge $U'_1$ of $\mathcal{H}'$ such that $U'_1 \subseteq G$. This would imply $U'_1 \subset U'$, contradicting the fact that $\mathcal{H}'$ is a simple hypergraph. □

**Corollary 3.87.** *Let $\mathcal{H} = (V, \mathcal{E})$ be a hypergraph and let $\mathcal{H}' = MINTR(\mathcal{H})$. Then, $\mathcal{H} = MINTR(\mathcal{H}')$.*

*Proof.* The corollary follows immediately by observing the symmetry of the conditions of Theorem 3.86 relative to the hypergraphs $\mathcal{H}$ and $\mathcal{H}'$. □

Corollary 3.87 can be written succinctly as

$$MINTR(MINTR(\mathcal{H})) = \mathcal{H} \tag{3.4}$$

for any simple hypergraph $\mathcal{H}$.

## Exercises and Supplements

1. How many graphs exist that have a finite set of vertices $V$ such that $|V| = n$?
2. Let $\mathcal{G} = (V, E)$ be a finite graph. How many spanning subgraphs exist for $\mathcal{G}$?
3. In Definition 3.11 it is required that the length $n$ of a cycle $\mathbf{p} = (v_0, \ldots, v_n)$ in a graph $\mathcal{G}$ be at least 3. Why is this necessary?
4. Let $S$ be a finite set. Define the graph $\mathcal{G}_S = (\mathcal{P}(S), E)$ such that an edge $(K, L)$ exists if $K \subset L$ and $|L| = |K|+1$. Prove that there exist $(|M|-|K|)!$ paths that join the vertex $K$ to $M$.
5. Let $\mathcal{G} = (V, E)$ be a finite graph. Define the triangular number of $\mathcal{G}$ as $t = \max\{|\Gamma(x) \cap \Gamma(y)| \mid (x, y) \in E\}$.
   a) Prove that for every two vertices $x, y \in V$ we have $\mathsf{d}(x)+\mathsf{d}(y) \leq |V|+t$.
   b) Show that
   $$\sum\{\mathsf{d}(x) + \mathsf{d}(y) \mid (x, y) \in E\} = \sum_{x \in V} \mathsf{d}^2(x).$$
   Conclude that
   $$\sum_{x \in V} \mathsf{d}^2(x) \leq (|V| + t)|E|.$$
6. Let $m(\mathcal{G})$ be the minimum degree of a vertex of the graph $\mathcal{G}$. Prove that $\mathcal{G}$ contains a path of length $m(\mathcal{G})$ and a cycle of length at least $m(\mathcal{G}) + 1$. **Hint:** Consider a path of maximal length in $\mathcal{G}$.
7. Let $\mathcal{C}$ be a collection of sets. The graph $\mathcal{G}_\mathcal{C}$ of $\mathcal{C}$ has $\mathcal{C}$ as the set of vertices. The set of edges consists of those pairs $C, D \in \mathcal{C}$ such that $C \neq D$ and $C \cap D \neq \emptyset$.
   a) Prove that for every graph $\mathcal{G}$ there exists a collection of sets $\mathcal{C}$ such that $\mathcal{G} = \mathcal{G}_\mathcal{C}$.
   b) Let $\mathcal{C} = (V, E)$ and let
   $$c(\mathcal{G}) = \min\{|S| \mid \mathcal{G} = \mathcal{G}_\mathcal{C} \text{ for some } \mathcal{C} \subseteq \mathcal{P}(S)\}.$$
   Prove that if $\mathcal{G}$ is connected and $|\mathcal{C}| \geq 3$, then $c(\mathcal{C}) \leq |E|$.

8. Let $\mathcal{G} = (V, E)$ be a graph such that $(u, v) \in (V \times V) - E$ implies $\mathsf{d}(u) + \mathsf{d}(v) \geq |V|$, where $V$ contains at least three vertices. Prove that $\mathcal{G}$ is connected.

   **Solution:** Suppose that $u$ and $v$ belong to two distinct connected components $C$ and $C'$ of the graph $\mathcal{G}$, where $|C| = p$ and $|C'| = q$, so $p+q \leq |V|$. If $x \in C$ and $y \in C'$, then there is no edge $(x, y)$, $\mathsf{d}(x) \leq p-1$ and $\mathsf{d}(y) \leq q-1$. Thus, $\mathsf{d}(x)+\mathsf{d}(y) \leq p+q-2 \leq |V|-2$, which contradicts the hypothesis. Thus, $u, v$ must belong to the same connected component, so $\mathcal{G}$ is connected.

Let $\mathcal{G} = (V, E)$ be a graph and let $C$ be a set referred to as the *set of colors*. A *$C$-coloring* of $\mathcal{G}$ is a function $f : V \longrightarrow C$ such that $(u, v) \in E$ implies $f(u) \neq f(v)$. The *chromatic number* of $\mathcal{G}$ is the least number $|C|$ such that the graph has a $C$-coloring. The chromatic number of $\mathcal{G}$ is denoted by $\chi(\mathcal{G})$; if $\chi(\mathcal{G}) = n$, then we say that $\mathcal{G}$ is *$n$-chromatic*.

9. Prove that the graph $\mathcal{G}$ has $\chi(\mathcal{G}) = 1$ if and only if it is totally disconnected. Further, prove that $\chi(\mathcal{G}) = 2$ if and only if it has no odd cycles.

Let $\mathcal{G} = (V, E)$ be a graph. A *Hamiltonian path* that joins the vertex $u$ to the vertex $v$ is a simple path in $\mathcal{G}$ that joins $u$ to $v$ such that every vertex of $V$ occurs in the path. A *Hamiltonian cycle* in $\mathcal{G}$ is a simple cycle that contains all vertices of $\mathcal{G}$. Next, we present several sufficient conditions for the existence of a Hamiltonian path due to O. Ore and G. A. Dirac.

10. Let $\mathcal{G} = (V, E)$ be a graph such that $(u, v) \in (V \times V) - E$ implies $\mathsf{d}(u) + \mathsf{d}(v) \geq |V|$, where $V$ contains at least three vertices. Prove that $\mathcal{G}$ contains a Hamiltonian cycle.

    **Solution:** The graph $\mathcal{G}$ is connected. Let $\mathbf{p} = (v_1, \ldots, v_m)$ be the longest simple path in $\mathcal{G}$.

    Suppose that $m = |V|$, which implies that $\mathbf{p}$ is a Hamiltonian path that joins $v_1$ to $v_m$. If $v_m$ and $v_1$ are joined by an edge, then $(v_1, \ldots, v_m, v_1)$ is a Hamiltonian cycle.

    Suppose that no edge exists between $v_m$ and $v_1$, so $\mathsf{d}(v_1) + \mathsf{d}(v_m) \geq |V|$. The vertex $v_1$ is joined to $\mathsf{d}(v_1)$ vertices on the path $(v_2, \ldots, v_m)$ and there are $\mathsf{d}(v_1)$ nodes that precede these nodes on the path $\mathbf{p}$. If $v_m$ were not joined to any of these nodes, then the set of nodes on $\mathbf{p}$ joined to $v_m$ would come from a set of $m - 1 - \mathsf{d}(v_1)$ nodes, so we would have $\mathsf{d}(v_m) \leq m - 1 - \mathsf{d}(v_1)$, which would contradict the assumption that $\mathsf{d}(v_1) + \mathsf{d}(v_m) \geq |V|$. Thus, there exists a node $v_i$ on $\mathbf{p}$ such that $(v_1, v_i), (v_{i-1}, v_n) \in E$. Therefore $(v_1, v_i, v_{i+1}, \ldots, v_n, v_{i-1}, v_{i-2}, \ldots, v_2, v_1$ is a Hamiltonian cycle.

    Suppose that $m < |V|$. If there is an edge $(w, v_1)$, where $w$ is not a node of $\mathbf{p}$, then $(w, v_1, \ldots, v_m)$ is longer than $\mathbf{p}$. Thus, $v_1$ is joined only to nodes in $\mathbf{p}$ and so is $v_m$. An argument similar to the one previously

used shows that there exists a simple cycle $\mathbf{q} = (y_1, \ldots, y_m, y_1)$ of length $m$. Since $m < |V|$ and $\mathcal{G}$ is connected, there is a node $t$ not on $\mathbf{q}$ that is joined to some vertex $y_k$ in $\mathbf{q}$. Then $(t, y_k, y_{k+1}, \ldots, y_m, y_1, y_2, \ldots, y_{k-1})$ is a path of length $m+1$, which contradicts the maximality of the length of $\mathbf{p}$. Since this case cannot occur, it follows that $\mathcal{G}$ has a Hamiltonian cycle.

11. Let $\mathcal{G} = (V, E)$ be a graph such that $|V| \geq 3$ in which $\mathsf{d}(v) \geq |V|/2$.
    a) Prove that $\mathcal{G}$ is connected.
    b) Prove that $\mathcal{G}$ has a Hamiltonian cycle.
12. Let $\mathcal{G} = (V, E)$ be a graph such that $|E| \geq \frac{(|V|-1)(|V|-2)}{2} + 2$. Prove that $\mathcal{G}$ has a Hamiltonian cycle.
13. Prove that a tree that has at least two vertices has at least two leaves.
14. Prove that if $\mathcal{T}$ is a tree and $x$ is a leaf, then the graph obtained by removing $x$ and the edge that has $x$ as an endpoint from $\mathcal{T}$ is again a tree. Also, show that if $z$ is a new vertex, then the graph obtained from $\mathcal{T}$ by joining $z$ with any node of $\mathcal{T}$ is also a tree.
15. Let $f : S \longrightarrow S$ be a function. Define the directed graph $\mathcal{G}_f = (S, E)$, where $E = \{(x, y) \in S \times S \mid y = f(x)\}$. Prove that each connected component of $\mathcal{G}_f$ consists of a cycle and a number of trees linked to the cycle.
16. Let $\mathcal{N} = (\mathcal{G}, cap, s, t)$ be a network. Prove that if $(C_1, C_1')$ and $(C_2, C_2')$ are minimal cuts, then $(C_1 \cup C_2, C_1' \cap C_2')$ is a minimal cut.
17. Let $\mathbf{M} \in \{0, 1\}^{m \times n}$ be a binary matrix. A *line* of $\mathbf{M}$ is either a row or a column; two 1s are independent if they are neither in the same row nor in the same column. A *line cover* of $\mathbf{M}$ is a set of lines such that every 1 is located in one of these lines. Prove that the maximum number of independent 1s equals the minimum size of a line cover.

    **Solution:** Let $R = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ be the set of rows of the matrix $\mathbf{M}$ and let $C = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be the set of columns of $\mathbf{M} = (m_{ij})$. Consider the bipartite graph $\mathcal{G}_{\mathbf{M}} = (R \cup C, E)$, where an edge exists in between the row $\mathbf{u}_i$ and the column $\mathbf{v}_j$ if and only if $m_{ij} = 1$. Two independent 1s correspond to two edges that have no common endpoints; therefore the maximum number of 1s equals the size of a maximal matching. The statement follows immediately from König's theorem for bipartite graphs (Theorem 3.80).
18. A hypergraph $\mathcal{H} = (V, \mathcal{E})$ is said to be $k$-uniform if $\mathcal{E} \subseteq \mathcal{P}_k(V)$. For a $k$-uniform hypergraph, where $k \geq 2$, prove that if $|V| \leq 2k$, then any two hyperedges have a nonempty intersection.
19. Let $\mathcal{H} = (V, \mathcal{E})$ be a hypergraph, where $|V| = n$. Prove that if $|\mathcal{E}| \leq |V|$, then there exists $v \in V$ such that for any hyperedges $U, U' \in \mathcal{E}$ we have $U - \{v\} \neq U' - \{v\}$.

    **Solution:** If $U$ and $U'$ are two distinct hyperedges, and $U - \{v\} = U' - \{v\}$, then $U \oplus U' = \{v\}$. Suppose that there is no vertex $v$ with the stated property. This amounts to saying that for every vertex $v \in V$ there exists a pair of hyperedges $U_{a(v)}$ and $U_{b(v)}$ such that $U_{a(v)} \oplus U_{b_v} = \{v\}$.

Let $\mathcal{G} = (\mathcal{E}, E)$ be a graph having the set of hyperedges of $\mathcal{H}$ as its vertices. An edge $(U, U')$ exists in this graph if and only if $U \oplus U' = \{v\}$. From the supposition made above, it follows that $|E| \geq |V|$.

We claim that $\mathcal{G}$ is acyclic. Suppose that $(U_1, \ldots, U_p, U_1)$ is a cycle of minimal length in $\mathcal{G}$. Then, there is a sequence of vertices in $\mathbf{Seq}(V)$ $(v_1, \ldots, v_p)$ such that $v_i \in U_i \oplus U_{i+1}$ for $1 \leq i \leq p-1$ and $v_p \in U_p \oplus U_1$. This implies

$$\begin{aligned}\{v_p\} &= U_1 \oplus U_p = (U_1 \oplus U_2) \oplus (U_2 \oplus U_3) \oplus \cdots \oplus (U_{p-1} \oplus U_p) \\ &\subseteq \bigcup_{i=1}^{p-1} (U_i \oplus U_{i+1}) = \{v_1, \ldots, v_{p-1}\},\end{aligned}$$

so $v_p = v_\ell$ for some $\ell$, $1 \leq \ell \leq p-1$. This contradicts the minimality of the cycle, so $\mathcal{G}$ is acyclic. However, this also leads to a contradiction because the acyclicity of $\mathcal{G}$ implies $|E| \leq |\mathcal{E}| - 1 \leq |V| - 1$.

20. Many concepts from graph theory have natural extensions to hypergraphs. For instance, the degree of a vertex $x$ in a hypergraph $\mathcal{H} = (V, \mathcal{E})$ is $d(x) = |\{U \in \mathcal{E} \mid x \in U\}|$.
Prove that there exists a hypergraph $\mathcal{H} = (V, \mathcal{E})$, where $V = \{v_1, \ldots, v_n\}$ and $\mathcal{E} = \{U_1, \ldots, U_m\}$ such that $d(x_i) = d_i$ for $1 \leq i \leq n$ and $|U_j| = r_j$ for $1 \leq j \leq m$, where $d_1 \geq d_2 \geq d_n$ if and only if the following conditions are satisfied:
a) $\sum_{j=1}^{m} \min\{r_j, k\} \geq d_1 + \cdots + d_k$ for $1 \leq k < n$.
b) $\sum_{j=1}^{m} r_j = d_1 + \cdots + d_n$.
**Hint:** Consider a network $\mathcal{N} = (\mathcal{G}, cap, s, t)$, where

$$\mathcal{G} = (\{s, x_1, \ldots, x_n, U_1, \ldots, U_m\}, E)$$

and the edges and their capacities are given by
- For every $U_j$, there is an edge $(s, U_j)$ with $cap(s, U_j) = r_j$.
- For every $x_i$, there is an edge $(x_i, t)$ with $cap(x_i, t) = d_i$.
- For every $U_j$ and $x_i$, there is an arc $(U_j, x_i)$ with $cap(U_j, x_i) = 1$.

Prove that a maximal flow in this network corresponds to a hypergraph that satisfies the desired conditions.

21. Let $\mathcal{G} = (V, E)$ be a digraph. If $U \subseteq V$, let $C(U) = \{(x, y) \in E \mid x \in U, y \notin U\}$ and let $f : \mathcal{P}(V) \longrightarrow \mathbb{R}_{\geq 0}$ be the function given by $f(U) = |C(U)|$ for $U \in \mathcal{P}(V)$. Prove that $f$ satisfies the inequality

$$f(U_1 \cup U_2) + f(U_1 \cap U_2) \leq f(U_1) + f(U_2)$$

for every $U_1, U_2 \in \mathcal{P}(V)$.

## Bibliographical Comments

Among the many references for graph theory we mention [65, 10, 18] and [38]. Theorem 3.86 is the "vertex-coloring lemma" from [9]. We adopted the

treatment of maximal flows (including Theorem 3.73) from [133], a classic reference for networks.

Supplement 10 is a result of O. Ore [102]; the same author (see [103]) has shown the statement in Exercise 12.

The result contained in Exercise 11 is from G. A. Dirac [39].

# Part II

# Partial Orders

# 4

# Partially Ordered Sets

## 4.1 Introduction

Partially ordered sets are mathematical structures that play a fundamental role in both mathematics and computer science. After introducing the notion of a partially ordered set (poset) and defining several classes of special elements associated with partial orders, we discuss closure and interior systems, topics that have multiple applications in topology, algebra, and data mining. Two partially ordered sets receive special attention: the poset of real numbers and the poset of partitions of a finite set.

Finally, partially ordered sets serve as the starting point for the study of several algebraic structures in Chapter 5.

## 4.2 Partial Orders

The fundamental notion of this chapter is introduced next.

**Definition 4.1.** *A* partial order on a set $S$ *is a relation* $\rho \subseteq S \times S$ *that is reflexive, antisymmetric, and transitive.*

*The pair* $(S, \rho)$ *is referred to as a* partially ordered set *or, for short, a* poset.

When $|S|$ is finite, we refer to poset $(S, \rho)$ as a *finite poset*.

A *strict partial order*, or more simply, *a strict order* on $S$, is a relation $\rho \subseteq S \times S$ that is irreflexive and transitive.

*Example 4.2.* The identity relation on a set $S$, $\iota_S$, is a partial order; this is often referred to as the *discrete partial order* on $S$. Also, the relation $\theta_S = S \times S$ is a partial order on $S$.

*Example 4.3.* The relation "$\leq$" on the set of partitions of a set $PART(S)$ introduced in Definition 1.113 is a partial order on the set $PART(S)$.

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_4, © Springer-Verlag London Limited 2008

*Example 4.4.* The relation $\delta$ introduced in Example 1.29 is a partial order on $\mathbb{N}$ since, as we have shown in Example 1.45, $\delta$ is reflexive, antisymmetric, and transitive.

For a poset $(S, \rho)$, we prefer to use the *infix notation*; that is, write $s\rho t$ instead of $(s,t) \in \rho$. Moreover, various partial orders have their traditional notations, which we favor. For example, the relation $\delta$ introduced in Example 1.29 is usually denoted by $|$. Therefore, we write $m \mid n$ to denote that $(m,n) \in \delta$. This is the well-known *divisibility relation* on $\mathbb{N}$. Whenever practical, for generic partially ordered sets, we denote their partial order relation by $\leq$. Generic strict partial orders will be denoted by $<$.

*Example 4.5.* The inclusion relation $\subseteq$ is a partial order on the set of subsets $\mathcal{P}(S)$ of a set $S$. The reader can easily verify that "$\subseteq$" is reflexive, antisymmetric, and transitive.

*Example 4.6.* Let $S$ be a set and let $\leq_{pref}$ be the relation on $\mathbf{Seq}(S)$ defined by $\mathbf{u} \leq_{pref} \mathbf{v}$ if $\mathbf{u}$ is a prefix of $\mathbf{v}$. Clearly, $\mathbf{u} \leq_{pref} \mathbf{v}$ if and only if there exists $t \in \mathbf{Seq}(S)$ such that $\mathbf{v} = \mathbf{ut}$. It is immediate that "$\leq_{pref}$" is a reflexive relation.

Suppose that $\mathbf{u} \leq_{pref} \mathbf{v}$ and $\mathbf{v} \leq_{pref} \mathbf{u}$. There exist $\mathbf{t}, \mathbf{t}' \in \mathbf{Seq}(S)$ such that $\mathbf{v} = \mathbf{ut}$ and $\mathbf{u} = \mathbf{vt}'$, which implies $\mathbf{u} = \mathbf{utt}'$. Thus, $\mathbf{tt}' = \boldsymbol{\lambda}$, so $\mathbf{t} = \mathbf{t}' = \boldsymbol{\lambda}$, which allows us to infer that $\mathbf{u} = \mathbf{v}$. This shows that the relation "$\leq_{pref}$" is antisymmetric.

Finally, suppose that $\mathbf{u} \leq_{pref} \mathbf{v}$ and $\mathbf{v} \leq_{pref} \mathbf{w}$. We have $\mathbf{v} = \mathbf{ut}$ and $\mathbf{w} = \mathbf{vs}$ for some $\mathbf{s}, \mathbf{t} \in \mathbf{Seq}(S)$. This implies $\mathbf{w} = \mathbf{uts}$, which shows that $\mathbf{u} \leq_{pref} \mathbf{w}$. Thus, "$\leq_{pref}$" is indeed a partial order on $\mathbf{Seq}(S)$.

In a similar manner, it is possible to show that the relations

$$\leq_{suff} = \{(\mathbf{u}, \mathbf{v}) \in (\mathbf{Seq}(S))^2 \mid \mathbf{v} = \mathbf{tu} \text{ for some } \mathbf{t} \in \mathbf{Seq}(S)\},$$
$$\leq_{inf} = \{(\mathbf{u}, \mathbf{v}) \in (\mathbf{Seq}(S))^2 \mid \mathbf{v} = \mathbf{tut}' \text{ for some } \mathbf{t}, \mathbf{t}' \in \mathbf{Seq}(S)\},$$

are partial orders on $\mathbf{Seq}(S)$ (exercise!).

If $(S, \leq)$ is a poset and $T \subseteq S$, then $(T, \leq_T)$ is also a poset, where $\leq_T = \leq \cap (T \times T)$ is the *trace of* $\leq$ *on* $T$.

Every strict partial order is also asymmetric. Indeed, let $<$ be a strict partial order on $S$ and assume that $x < y$. If $y < x$, then $x < x$ due to the transitivity of $<$, which contradicts the irreflexivity of $<$. This shows that $<$ is indeed asymmetric.

A strict partial order is not, in general, a partial order since strict partial orders are irreflexive, while partial orders are reflexive. The link between partial orders and strict partial orders is given next.

**Theorem 4.7.** *If $\leq$ is a partial order on a set $S$ and $<\, = \leq - \iota_S$, then the relation $<$ is a strict partial order on $S$.*

*Conversely, if $<$ is a strict partial order on $S$, then $\leq = < \cup \iota_S$ is a partial order on $S$.*

*Proof.* Since $\iota_S \cap < = \emptyset$, the relation $<$ is irreflexive.

To prove the transitivity of $<$, let $x, y, z \in S$ be such that $x < y, y < z$. Because of the transitivity of $\leq$, we have $x \leq z$. On the other hand, we also have $x \neq z$. Indeed, if we assume that $x = z$, then we would have both $z < y$ and $y < z$, which is impossible by the asymmetry of $<$. Therefore, $(x, z) \in \leq - \iota_S = <$, which implies the transitivity of $<$.

Now, let $<$ be a strict partial order and let $\leq = < \cup \iota_S$. The reflexivity of $\leq$ is immediate.

To show that $\leq$ is antisymmetric, assume that $x \leq y$ and $y \leq x$. Because of the definition of $\leq$, we may have $x < y$ or $(x, y) \in \iota_S$ (that is, $x = y$). In the first case, we have a contradiction. Indeed, if $y < x$, this contradicts the asymmetry of $<$; if $(y, x) \in \iota_S$, we also have $(x, y) \in \iota_S$, and this contradicts the irreflexivity of $<$. Consequently, we must have $x = y$.

Let $x \leq y$ and $y \leq x$. We need to consider the following four cases.

(i) If $x < y, y < z$, we have $x < z$ because of the transitivity of $<$. This implies $x \leq z$.
(ii) If $(x, y) \in \iota_S$ and $y < z$, we have $x = y$; hence, $x < z$ and therefore $x \leq z$.
(iii) If $x < y$ and $(y, z) \in \iota_S$, we follow an argument similar to the one used in the previous case.
(iv) If $(x, y), (y, z) \in \iota_S$, we have $(x, z) \in \iota_S$ because of the transitivity of $\iota_S$; hence, $x \leq z$.

We proved that $\leq$ is also transitive, and this concludes our argument. □

*Example 4.8.* Consider the relation $\leq \subseteq \mathbb{R} \times \mathbb{R}$, which is a partial order. The strict partial order attached to it by the previous proposition is the relation "$<$".

A relation $\rho \subseteq S \times S$ is *acyclic* if $\rho^n \cap \iota_S = \emptyset$ for every $n \geq 1$.

Acyclicity is a hereditary property; this means that if a relation $\sigma \subseteq S \times S$ is acyclic and $\theta \subseteq \sigma$, then $\theta$ is also acyclic.

**Theorem 4.9.** *Every strict partial order is acyclic.*

*Proof.* Let $\rho$ be a strict partial order relation on $S$. Its transitivity implies the existence of the descending sequence

$$\cdots \subseteq \rho^n \subseteq \cdots \subseteq \rho^2 \subseteq \rho.$$

Since $\rho$ is irreflexive, we have $\rho \cap \iota_S = \emptyset$, and this implies $\rho^n \cap \iota_S = \emptyset$. □

Next we introduce a graphical representation of partial orders.

**Definition 4.10.** *Let $(S, \leq)$ be a poset. The* Hasse diagram *of $(S, \leq)$ is the digraph of the relation $< - (<)^2$, where "$<$" is the strict partial order corresponding to $\leq$.*

In view of the properties of acyclic relations discussed above, it is clear that the relation $< -(<)^2$ is acyclic; therefore, the Hasse diagram is always an acyclic directed graph. We will denote this relation by "$\prec$".

Observe that $x \prec y$ if $x \neq y$, $x \leq y$, and there is no $u \in S$ such that $x \leq u$ and $u \leq y$. In other words, if $x \prec y$, then $y$ *covers* $x$ directly, without any intermediate elements.

The use of Hasse diagrams in representing posets is justified by the following statement.

**Theorem 4.11.** *If $\leq$ is a partial order on a finite set $S$, $<$ is the strict partial order corresponding to $\leq$, and $\theta =< -(<)^2$, then $\theta^* =\leq$.*

*Proof.* Let $x, y \in S$ such that $x \leq y$. If $x = y$, then we have $(x, y) \in \iota_S \subseteq \theta^*$.

Assume now that $x \leq y$ and $x \neq y$, which means that $x < y$. Consider the collection $\mathcal{C}_{xy}$ of all sequences of elements of $A$ that can be "interpolated" between $x$ and $y$:

$$\mathcal{C}_{xy} = \{(\mathbf{s}(0), \ldots, \mathbf{s}(n-1)) \mid x = \mathbf{s}(0), \mathbf{s}(n-1) = y, \text{ and } \mathbf{s}(i) < \mathbf{s}(i+1) \text{ for } 0 \leq i \leq n-2, n \geq 2\}.$$

We have $\mathcal{C}_{xy} \neq \emptyset$ since the sequence $(x, y)$ belongs to $\mathcal{C}_{xy}$. Furthermore, no sequence from $\mathcal{C}_{xy}$ may contain a repetition because, if we have $s(p) = s(q)$ for a sequence $(\mathbf{s}(0), \ldots, \mathbf{s}(n-1))$ with $1 \leq p < q \leq n$, by the transitivity of $<$, this implies $\mathbf{s}(p) < \mathbf{s}(q)$. This is a contradiction because $< \cap \iota_S = \emptyset$. Since $S$ is finite, $\mathcal{C}_{xy}$ contains a finite number of sequences.

Consider a sequence of maximal length from $\mathcal{C}_{xy}$,

$$(\mathbf{s}(0), \mathbf{s}(2), \ldots, \mathbf{s}(m-1)),$$

where $x = \mathbf{s}(0)$ and $y = \mathbf{s}(m-1)$. Observe that for no pair $(\mathbf{s}(i), \mathbf{s}(i+1))$ can we have $(\mathbf{s}(i), \mathbf{s}(i+1)) \in (<)^2$. Indeed, if $(\mathbf{s}(j), \mathbf{s}(j+1)) \in (<)^2$, then there is $x \in S$ such that $\mathbf{s}(i) < x$ and $x < \mathbf{s}(i+1)$, and this contradicts the maximality of $m$. Therefore, $(\mathbf{s}(i), \mathbf{s}(i+1)) \in < -(<)^2 = \theta$, and this shows that $(x, y) \in \theta^{m-1} \subseteq \theta^*$.

Conversely, if $(x, y) \in \theta^*$, there is $k \in \mathbb{N}$ such that $(x, y) \in \theta^k$, which means that there exists a sequence $(\mathbf{z}(0), \ldots, \mathbf{z}(k))$ such that

$$x = \mathbf{z}(0), (\mathbf{z}(i), \mathbf{z}(i+1)) \in \theta \text{ for } 0 \leq i \leq k-1 \text{ and } y = \mathbf{z}(k).$$

This implies $\mathbf{z}(i) \leq \mathbf{z}(i+1)$; hence, $(x, y) \in (\leq)^k \subseteq \leq$ because of the transitivity of $\leq$. □

The relation $\theta$ introduced in Theorem 4.11 is called the *transitive reduction* of the partial order $\rho$.

*Example 4.12.* The Hasse diagram of the poset $(\mathcal{P}(S), \subseteq)$, where $S = \{a, b, c\}$, is given in Figure 4.1.

*Example 4.13.* Consider the poset $(\{2, 3, 4, 5, 6, 7, 8\}, \delta)$, where $\delta$ is the divisibility relation introduced in Example 1.29. Its Hasse diagram is shown in Figure 4.2.

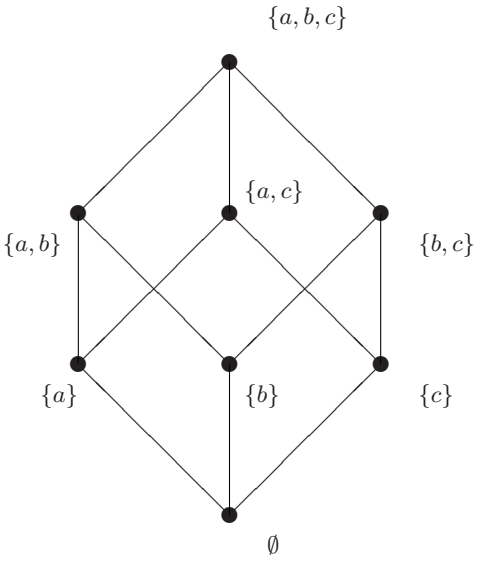



**Fig. 4.1.** Hasse diagram of the poset $(\mathcal{P}(S), \subseteq)$.

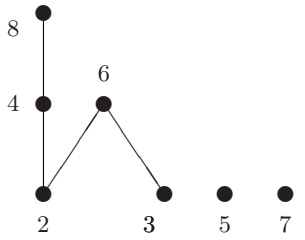



**Fig. 4.2.** Hasse diagram of the poset $(\{2, 3, 4, 5, 6, 7, 8\}, \delta)$.

## 4.3 Special Elements of Partially Ordered Sets

Let $(S, \leq)$ be a poset and let $K \subseteq S$.

**Definition 4.14.** *The* set of upper bounds *of the set $K$ is the set*

$$K^s = \{y \in S \mid x \leq y \text{ for every } x \in K\}.$$

*The* set of lower bounds *of the set $K$ is the set*

$$K^i = \{y \in S \mid y \leq x \text{ for every } x \in K\}.$$

*If $K^s \neq \emptyset$, we say that the set $K$ is* bounded above. *Similarly, if $K^i \neq \emptyset$, we say that $K$ is* bounded below. *If $K$ is both bounded above and bounded below we will refer to $K$ as a* bounded set.

*If $K^s = \emptyset$ ($K^i = \emptyset$), then $K$ is said to be* unbounded above (below).

**Theorem 4.15.** *Let $(S, \leq)$ be a poset and let $U$ and $V$ be two subsets of $S$. If $U \subseteq V$, then we have $V^i \subseteq U^i$ and $V^s \subseteq U^s$.*

*Also, for every subset $T$ of $S$, we have $T \subseteq (T^s)^i$ and $T \subseteq (T^i)^s$.*

*Proof.* The argument for both statements of the theorem amounts to a direct application of Definition 4.14. □

Note that for every subset $T$ of a poset $S$, we have both

$$T^i = ((T^i)^s)^i \tag{4.1}$$

and

$$T^s = ((T^s)^i)^s. \tag{4.2}$$

Indeed, since $T \subseteq (T^i)^s$, by the first part of Theorem 4.15, we have $((T^s)^i)^s \subseteq T^s$. By the second part of the same theorem applied to $T^s$, we have the reverse inclusion $T^s \subseteq ((T^s)^i)^s$, which yields $T^s = ((T^s)^i)^s$.

**Theorem 4.16.** *For any subset $K$ of a poset $(S, \rho)$, the sets $K \cap K^s$ and $K \cap K^i$ contain at most one element.*

*Proof.* Suppose that $y_1, y_2 \in K \cap K^s$. Since $y_1 \in K$ and $y_2 \in K^s$, we have $(y_1, y_2) \in \rho$. Reversing the roles of $y_1$ and $y_2$ (that is, considering now that $y_2 \in K$ and $y_1 \in K^s$), we obtain $(y_2, y_1) \in \rho$. Therefore, we may conclude that $y_1 = y_2$ because of the antisymmetry of the relation $\rho$, which shows that $K \cap K^s$ contains at most one element.

A similar argument can be used for the second part of the proposition; we leave it to the reader. □

**Definition 4.17.** *Let $(S, \leq)$ be a poset. The* least (greatest) element *of the subset $K$ of $S$ is the unique element of the set $K \cap K^i$ ($K \cap K^s$, respectively) if such an element exists.*

If $K$ is unbounded above, then it is clear that $K$ has no greatest element. Similarly, if $K$ is unbounded below, then $K$ has no least element.

Applying Definition 4.17 to the set $S$, the *least (greatest) element* of the poset $(S, \leq)$ is an element $a$ of $S$ such that $a \leq x$ ($x \leq a$, respectively) for all $x \in S$.

It is clear that if a poset has a least element $u$, then $u$ is the unique minimal element of that poset. A similar statement holds for the greatest and the maximal elements.

**Definition 4.18.** *Let $(S, \leq)$ be a poset that has 0 as its least element. An* atom *of $(S, \leq)$ is an element $x$ of $S$ such that $0 \prec x$.*

*If $(S, \leq)$ is a poset that has 1 as its greatest element, then $y$ is a* co-atom *of $(S, \leq)$ if $y \leq 1$.*

*Example 4.19.* For the poset introduced in Example 4.12, the greatest element is $\{a, b, c\}$, while the least element is $\emptyset$.

The atoms of this poset are $\{a\}, \{b\}, \{c\}$; its co-atoms are $\{a, b\}, \{b, c\}$, and $\{a, c\}$.

**Definition 4.20.** *The subset $K$ of the poset $(S, \leq)$ has a* least upper bound *$u$ if $K^s \cap (K^s)^i = \{u\}$.*

*$K$ has the* greatest lower bound *$v$ if $K^i \cap (K^i)^s = \{v\}$.*

We note that a set can have at most one least upper bound and at most one greatest lower bound. Indeed, we have seen above that for any set $U$ the set $U \cap U^i$ may contain an element or be empty. Applying this remark to the set $K^s$, it follows that the set $K^s \cap (K^s)^i$ may contain at most one element, which shows that $K$ may have at most one least upper bound. A similar argument can be made for the greatest lower bound.

If the set $K$ has a least upper bound, we denote it by $\sup K$. The greatest lower bound of a set will be denoted by $\inf K$. These notations come from the terms *supremum* and *infimum* used alternatively for the least upper bound and the greatest lower bound, respectively.

*Example 4.21.* Consider the poset $(\mathbb{N}, \delta)$, and let $m$ and $n$ be two distinct natural numbers $m \neq n$.

We claim that any set $\{m, n\}$ has both an infimum and a supremum. Indeed, let $p$ be the least common multiple of $m$ and $n$. Since $(n, p), (m, p) \in \delta$, it is clear that $p$ is an upper bound of the set $\{m, n\}$. On the other hand, if $k$ is an upper bound of $\{m, n\}$, then $k$ is a multiple of both $m$ and $n$. In this case, $k$ must also be a multiple of $p$ because otherwise we could write $k = pq + r$ with $0 < r < p$ by dividing $k$ by $p$. This would imply $r = k - pq$; hence, $r$ would be a multiple of both $m$ and $n$ because both $k$ and $p$ have this property. However, this would contradict the fact that $p$ is the least multiple that $m$ and $n$ share! This shows that the least common multiple of $m$ and $n$ coincides with the supremum of the set $\{m, n\}$.

The reader can easily prove that the infimum of $\{m, n\}$ coincides with the greatest common divisor of the numbers $m$ and $n$.

*Example 4.22.* Consider a set $M$ and the poset $(\mathcal{P}(M), \subseteq)$. Let $K$ and $H$ be two subsets of $M$. The set $\{K, H\}$ has an infimum and a supremum. Indeed, let $L = K \cap H$. Clearly, $L \subseteq K$ and $L \subseteq H$, so $L$ is a lower bound of the set $\{K, H\}$. Furthermore, if $J \subseteq K$ and $J \subseteq H$, then $J \subseteq L$ by the definition of the intersection. This proves that the infimum of $\{K, H\}$ is the intersection $K \cap H$. A similar argument shows that $K \cup H$ is the supremum of $\{K, H\}$.

In the previous two examples, any two-element subset of the poset has both a supremum and an infimum.

For a one-element subset $\{x\}$ of a poset $(S, \rho)$, we have $\sup\{x\} = \inf\{x\} = x$.

**Definition 4.23.** *A* minimal element *of a poset $(S, \leq)$ is an element $x \in S$ such that $\{x\}^i = \{x\}$. A* maximal element *of $(S, \leq)$ is an element $y \in S$ such that $\{y\}^s = \{y\}$.*

In other words, $x$ is a minimal element of the poset $(S, \leq)$ if there is no element less than or equal to $x$ other than itself; similarly, $x$ is maximal if there is no element greater than or equal to $x$ other than itself.

The set of minimal elements of a poset $(S, \leq)$ is denoted by $MIN(S, \leq)$; the set of maximal elements of this poset is denoted by $MAX(S, \leq)$.

*Example 4.24.* Not every subset of a poset has a least or a greatest element. Indeed, let $(\{2, 3, 4, 5, 6, 7, 8, \}, \delta)$ be a poset whose Hasse diagram is shown in Figure 4.2. It is easy to see that

$$MIN(\{2, 3, 4, 5, 6, 7, 8, \}, \delta) = \{2, 3, 5, 7\},$$
$$MAX(\{2, 3, 4, 5, 6, 7, 8, \}, \delta) = \{5, 6, 7, 8\}.$$

There is no least element and there is no largest element in this poset.

**Theorem 4.25.** *Every finite nonempty subset $K$ of a poset $(S, \leq)$ has a minimal element and a maximal element.*

*Proof.* Suppose that $K = \{x_0, \ldots, x_{n-1}\}$ for $n \geq 1$. Define the element $u_0 = x_0$ and

$$u_k = \begin{cases} x_k & \text{if } x_k < u_{k-1}, \\ u_{k-1} & \text{otherwise.} \end{cases}$$

Then, $u_{n-1}$ is a minimal element. The proof of the existence of a maximal element of $K$ is similar. □

Next, we discuss a simple property of partially ordered sets that will allow us to obtain half of some of the arguments related to the properties of partial orders for free.

**Theorem 4.26.** *Let $\rho$ be a partial order on a set $S$. The inverse $\rho^{-1}$ is also a partial order on the same set.*

*Proof.* Since $(x, x) \in \rho$ for every $x \in S$, it follows that $(x, x) \in \rho^{-1}$ for every $x \in S$, so $\rho^{-1}$ is reflexive.

The antisymmetry of $\rho^{-1}$ follows from $(\rho^{-1})^{-1} = \rho$ and because of the antisymmetry of $\rho$.

To prove the transitivity of $\rho^{-1}$, assume that $(x, y) \in \rho^{-1}$ and $(y, z) \in \rho^{-1}$. This means that $(y, x), (z, y) \in \rho$, and because of the transitivity of $\rho$, we obtain $(z, x) \in \rho$, so $(x, z) \in \rho^{-1}$, which proves that $\rho^{-1}$ is transitive. □

**Definition 4.27.** *The* dual *of the poset* $(S, \rho)$ *is the poset* $(S, \rho^{-1})$.

Concepts valid for a poset have a counterpart for their dual poset. For instance, $x$ is an upper bound for the set $K$ in the poset $(S, \rho)$ if and only if $x$ is a lower bound for $K$ in the dual poset. Similarly, $t = \sup K$ in the poset $(S, \rho)$ if and only if $t = \inf K$ in the dual poset. Similar pairs are minimal element and maximal element, infimum and supremum, etc.

If all concepts occurring in a statement about posets are replaced by their duals, we obtain the *dual statement*; the method of proving statements about posets is known as *dualization*. Furthermore, if a statement holds for a poset $(S, \rho)$, its dual holds for the dual poset $(S, \rho^{-1})$. This allows us to formulate the following principle.

**The Duality Principle for Posets:** If a statement is true for all posets, then its dual is also true for all posets.

The validity of this principle follows from the fact that any poset can be regarded as the dual of some other poset. The duality principle allows us to simplify proofs of certain statements that concern posets. For statements involving both a concept and its dual we need to prove only half of the statement; the other half follows by applying the duality principle. For instance, once we prove the statement "any subset of a poset can have at most one least upper bound," the dual statement "any subset of a poset can have at most one greatest lower bound" follows.

## 4.4 The Poset of Real Numbers

For the poset $(\mathbb{R}, \leq)$, it is possible to give more specific descriptions of the supremum and infimum of a subset when they exist.

**Theorem 4.28.** *If* $T \subseteq \mathbb{R}$, *then* $u = \sup T$ *if and only if* $u$ *is an upper bound of* $T$ *and, for every* $\epsilon > 0$, *there is* $t \in T$ *such that* $u - \epsilon < t \leq u$.

*The number* $v$ *is* $\inf T$ *if and only if* $v$ *is a lower bound of* $T$ *and, for every* $\epsilon > 0$, *there is* $t \in T$ *such that* $v \leq t < v + \epsilon$.

*Proof.* We prove only the first part of the theorem; the argument for the second part is similar and is left to the reader.

Suppose that $u = \sup T$; that is, $\{u\} = T^s \cup (T^s)^i$. Since $u \in T^s$, it is clear that $u$ is an upper bound for $T$. Suppose that there is $\epsilon > 0$ such that no $t \in T$ exists such that $u - \epsilon < t \leq u$. This means that $u - \epsilon$ is also an upper bound for $T$, and in this case $u$ cannot be a lower bound for the set of upper bounds of $T$. Therefore, no such $\epsilon$ may exist.

Conversely, suppose that $u$ is an upper bound of $T$ and for every $\epsilon > 0$, there is $t \in T$ such that $u - \epsilon < t \leq u$. Suppose that $u$ does not belong to $(K^s)^i$. This means that there is another upper bound $u'$ of $T$ such that $u' < u$. Choosing $\epsilon = u - u'$, we would have no $t \in T$ such that $u - \epsilon = u' < t \leq u$

because this would prevent $u'$ from being an upper bound of $T$. This implies $u \in (K^s)^i$, so $u = \sup T$. □

A very important axiom for the set $\mathbb{R}$ is given next.

**The Completeness Axiom for $\mathbb{R}$:** If $T$ is a nonempty subset of $\mathbb{R}$ that is bounded above, then $T$ has a supremum.

A statement equivalent to the Completeness Axiom for $\mathbb{R}$ follows.

**Theorem 4.29.** *If $T$ is a nonempty subset of $\mathbb{R}$ that is bounded below, then $T$ has an infimum.*

*Proof.* Note that the set $T^i$ is not empty. If $s \in T^i$ and $t \in T$, we have $s \leq t$, so the set $T^i$ is bounded above. By the Completeness axiom $v = \sup T^i$ exists and $\{v\} = (T^i)^s \cap ((T^i)^s)^i = (T^i)^s \cap T^i$ by Equality (4.1). Thus, $v = \inf T$. □

We leave to the reader to prove that Theorem 4.29 implies the Completeness Axiom for $\mathbb{R}$.

Another statement equivalent to the Completeness Axiom is the following.

**Theorem 4.30 (Dedekind's Theorem).** *Let $U$ and $V$ be nonempty subsets of $\mathbb{R}$ such that $U \cup V = \mathbb{R}$ and $x \in U, y \in V$ imply $x < y$. Then, there exists $a \in \mathbb{R}$ such that if $x > a$, then $x \in V$, and if $x < a$, then $x \in U$.*

*Proof.* Observe that $U \neq \emptyset$ and $V \subseteq U^s$. Since $V \neq \emptyset$, it means that $U$ is bounded above, so by the Completeness Axiom $\sup U$ exists. Let $a = \sup U$. Clearly, $u \leq a$ for every $u \in U$. Since $V \subseteq U^s$, it also follows that $a \leq v$ for every $v \in V$.

If $x > a$, then $x \in V$ because otherwise we would have $x \in U$ since $U \cup V = \mathbb{R}$ and this would imply $x \leq a$. Similarly, if $x < a$, then $x \in U$. □

Using the previously introduced notations, Dedekind's theorem can be stated as follows: if $U$ and $V$ are nonempty subsets of $\mathbb{R}$ such that $U \cup V = \mathbb{R}$, $U^s \subseteq V$, $V^i \subseteq U$, then there exists $a$ such that $\{a\}^s \subseteq V$ and $\{a\}^i \subseteq U$.

One can prove that Dedekind's theorem implies the Completeness Axiom. Indeed, let $T$ be a nonempty subset of $\mathbb{R}$ that is bounded above. Therefore $V = T^s \neq \emptyset$. Note that $U = (T^s)^i \neq \emptyset$ and $U \cup V = \mathbb{R}$. Moreover, $U^s = ((T^s)^i)^s = T^s = V$ and $V^i = (T^s)^i = U$. Therefore, by Dedekind's theorem, there is $a \in \mathbb{R}$ such that $\{a\}^s \subseteq V = T^s$ and $\{a\}^i \subseteq U = (T^s)^i$. Note that $a \in \{a\}^s \cap \{a\}^i \subseteq T^s \cap (T^s)^i$, which proves that $a = \sup T$.

By adding the symbols $+\infty$ and $-\infty$ to the set $\mathbb{R}$, one obtains the set $\hat{\mathbb{R}}$. The partial order $\leq$ defined on $\mathbb{R}$ can now be extended to $\hat{\mathbb{R}}$ by $-\infty \leq x$ and $x \leq +\infty$ for every $x \in \mathbb{R}$.

We also extend the addition and multiplication of reals to $\hat{\mathbb{R}}$ by

$$
\begin{aligned}
x + \infty = +\infty + x &= +\infty \text{ for } -\infty < x \leq +\infty,\\
x - \infty = -\infty + x &= -\infty \text{ for } -\infty \leq x < +\infty,\\
x \cdot \infty = \infty \cdot x \quad &= \begin{cases} -\infty & \text{if } -\infty \leq x < 0,\\ 0 & \text{if } x = 0\\ \infty & \text{if } 0 < x \leq +\infty \end{cases},\\
x \cdot (-\infty) = -\infty \cdot x \quad &= \begin{cases} \infty & \text{if } -\infty \leq x < 0,\\ 0 & \text{if } x = 0,\\ -\infty & \text{if } 0 < x \leq +\infty \end{cases},\\
\tfrac{x}{+\infty} = \tfrac{x}{-\infty} \quad &= 0 \text{ for } x \in \mathbb{R}.
\end{aligned}
$$

The operations $+\infty - \infty$ and $-\infty + \infty$ are undefined.

Note that, in the poset $(\hat{\mathbb{R}}, \leq)$, the sets $T^i$ and $T^s$ are nonempty for every $T \in \mathcal{P}(\hat{\mathbb{R}})$ because $-\infty \in T^i$ and $+\infty \in T^s$ for any subset $T$ of $\hat{\mathbb{R}}$.

**Theorem 4.31.** *For every set $T \subseteq \hat{\mathbb{R}}$, both $\sup T$ and $\inf T$ exist in the poset $(\hat{\mathbb{R}}, \leq)$.*

*Proof.* We present the argument for $\sup T$. If $\sup T$ exists in $(\mathbb{R}, \leq)$, then it is clear that the same number is $\sup T$ in $(\hat{\mathbb{R}}, \leq)$.

Assume now that $\sup T$ does not exist in $(\mathbb{R}, \leq)$. By the Completeness Axiom for $\mathbb{R}$, this means that the set $T$ does not have an upper bound in $(\mathbb{R}, \leq)$. Therefore, the set of upper bounds of $T$ in $(\hat{T}, \leq)$ is $T^{\hat{s}} = \{+\infty\}$. It follows immediately that in this case $\sup T = +\infty$ in $(\hat{\mathbb{R}}, \leq)$. □

## 4.5 Closure and Interior Systems

The notions of closure system and interior system introduced in this section are significant in algebra and topology and have applications in the study of frequent item sets in data mining.

**Definition 4.32.** *Let $S$ be a set. A* closure system on $S$ *is a collection $\mathcal{C}$ of subsets of $S$ that satisfies the following conditions:*
(i) $S \in \mathcal{C}$ *and*
(ii) *for every collection $\mathcal{D} \subseteq \mathcal{C}$, we have $\bigcap \mathcal{D} \in \mathcal{C}$.*

*Example 4.33.* Let $\mathcal{C}$ be the collection of all intervals $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$ with $a, b \in \mathbb{R}$ and $a \leq b$ together with the empty set and the set $\mathbb{R}$. Note that $\bigcup \mathcal{C} = \mathbb{R} \in \mathcal{C}$, so the first condition of Definition 4.32 is satisfied.

Let $\mathcal{D}$ be a nonempty subcollection of $\mathcal{C}$. If $\emptyset \in \mathcal{D}$, then $\bigcap \mathcal{D} = \emptyset \in \mathcal{C}$. If $\mathcal{D} = \{\mathbb{R}\}$, then $\bigcap \mathcal{D} = \mathbb{R} \in \mathcal{C}$. Therefore, we need to consider only the case when $\mathcal{D} = \{[a_i, b_i] \mid i \in I\}$. Then, $\bigcap \mathcal{D} = \emptyset$ unless $a = \sup\{a_i \mid i \in I\}$ and $b = \inf\{b_i \mid i \in I\}$ both exist and $a \leq b$, in which case $\bigcap \mathcal{D} = [a, b]$. Thus, $\mathcal{C}$ is a closure system.

*Example 4.34.* Let $\mathcal{A} = (A, \mathcal{I})$ be an algebra and let $\mathsf{S}(\mathcal{A})$ be the collection of subalgebras of $\mathcal{A}$, $\mathsf{S}(\mathcal{A}) = \{(A_i, \mathcal{I}) \mid i \in I\}$. The collection $\mathcal{S} = \{A_i \mid i \in I\}$ is a closure system. It is clear that we have $S \in \mathcal{S}$. Also, if $\{A_i \mid i \in J\}$ is a family of subalgebras, then $\bigcap_{i \in J} A_i$ is a subalgebra of $\mathcal{A}$.

Many classes of relations define useful closure systems.

**Theorem 4.35.** *Let $S$ be a set and let $REFL(S), SYMM(S)$ and $TRAN(S)$ be the sets of reflexive relations, the set of symmetric relations, and the set of transitive relations on $S$, respectively. Then, $REFL(S), SYMM(S)$ and $TRAN(S)$ are closure systems on $S$.*

*Proof.* Note that $S \times S$ is a reflexive, symmetric, and transitive relation on $S$. Therefore, $\bigcup REFL(S) = S \times S \in REFL(S)$, $\bigcup SYMM(S) = S \times S \in SYMM(S)$, and $\bigcup TRAN(S) = S \times S \in TRAN(S)$.

Now let $\mathcal{C} = \{\rho_i \mid i \in I\}$ be a collection of transitive relations and let $\rho = \bigcap\{\rho_i \mid i \in I\}$. Suppose that $(x, y), (y, z) \in \rho$. Then $(x, y), (y, z) \in \rho_i$ for every $i \in I$, so $(x, z) \in \rho_i$ for $i \in I$ because each of the relations $\rho_i$ is transitive. Thus, $(x, z) \in \rho$, which shows that $\bigcap \mathcal{C} \in TRAN(S)$. This allows us to conclude that $TRAN(S)$ is indeed a closure system. We leave it to the reader to prove that $REFL(S)$ and $SYMM(S)$ are also closure systems. □

**Theorem 4.36.** *The set of equivalences on $S$, $EQS(S)$, is a closure system.*

*Proof.* The relation $\theta_S = S \times S$, is clearly an equivalence relation as we have seen in the proof of Theorem 4.35. Thus, $\bigcup EQS(S) = \theta_S \in EQS(S)$.

Now let $\mathcal{C} = \{\rho_i \mid i \in I\}$ be a collection of transitive relations and let $\rho = \bigcap\{\rho_i \mid i \in I\}$. It is immediate that $\rho$ is an equivalence on $S$, so $EQS(S)$ is a closure system. □

**Definition 4.37.** *A mapping $\boldsymbol{K} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ is a* closure operator *on a set $S$ if it satisfies the conditions*
(i) $U \subseteq \boldsymbol{K}(U)$ (expansiveness),
(ii) $U \subseteq V$ *implies* $\boldsymbol{K}(U) \subseteq \boldsymbol{K}(V)$ (monotonicity), *and*
(iii) $\boldsymbol{K}(\boldsymbol{K}(U)) = \boldsymbol{K}(U)$ (idempotency)
*for $U, V \in \mathcal{P}(S)$.*

*Example 4.38.* Let $\mathbf{K} : \mathcal{P}(\mathbb{R}) \longrightarrow \mathcal{P}(\mathbb{R})$ be defined by

$$\mathbf{K}(U) = \begin{cases} \emptyset & \text{if } U = \emptyset, \\ [a, b] & \text{if both } a = \inf U \text{ and } b = \sup U \text{ exist}, \\ \mathbb{R} & \text{otherwise}, \end{cases}$$

for $U \in \mathcal{P}(\mathbb{R})$. We leave to the reader the verification that $\mathbf{K}$ is a closure operator.

Closure operators induce closure systems, as shown by the next lemma.

**Lemma 4.39.** *Let $\boldsymbol{K} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ be a closure operator. Define the family of sets $\mathcal{C}_{\boldsymbol{K}} = \{H \in \mathcal{P}(S) \mid H = \boldsymbol{K}(H)\}$. Then, $\mathcal{C}_{\boldsymbol{K}}$ is a closure system on $S$.*

*Proof.* Since $S \subseteq \mathbf{K}(S) \subseteq S$, we have $S \in \mathcal{C}_{\mathbf{K}}$, so $\bigcup \mathcal{C}_{\mathbf{K}} = S \in \mathcal{C}_{\mathbf{K}}$.

Let $\mathcal{D} = \{D_i \mid i \in I\}$ be a collection of subsets of $S$ such that $D_i = \mathbf{K}(D_i)$ for $i \in I$. Since $\bigcap \mathcal{D} \subseteq D_i$, we have $\mathbf{K}(\bigcap \mathcal{D}) \subseteq \mathbf{K}(D_i) = D_i$ for every $i \in I$. Therefore, $\mathbf{K}(\bigcap \mathcal{D}) \subseteq \bigcap \mathcal{D}$, which implies $\mathbf{K}(\bigcap \mathcal{D}) = \bigcap \mathcal{D}$. This proves our claim. □

Note that $\mathcal{C}_{\mathbf{K}}$, as defined in Lemma 4.39, equals the range of $\mathbf{K}$. Indeed, if $L \in \mathrm{Ran}(\mathbf{K})$, then $L = \mathbf{K}(H)$ for some $H \in \mathcal{P}(S)$, so $\mathbf{K}(L) = \mathbf{K}(\mathbf{K}(H)) = \mathbf{K}(H) = L$, which shows that $L \in \mathcal{C}_{\mathbf{K}}$. The reverse inclusion is obvious.

We refer to the sets in $\mathcal{C}_{\mathbf{K}}$ as the *$\boldsymbol{K}$-closed subsets of $S$*.

In the reverse direction from Lemma 4.39, we show that every closure system generates a closure operator.

**Lemma 4.40.** *Let $\mathcal{C}$ be a closure system on the set $S$. Define the mapping $\boldsymbol{K}_{\mathcal{C}} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ by $\boldsymbol{K}_{\mathcal{C}}(H) = \bigcap\{L \in \mathcal{C} \mid H \subseteq L\}$. Then, $\boldsymbol{K}_{\mathcal{C}}$ is a closure operator on the set $S$.*

*Proof.* Note that the collection $\{L \in \mathcal{C} \mid H \subseteq L\}$ is not empty since it contains at least $S$, so $\mathbf{K}_{\mathcal{C}}(H)$ is defined and is clearly the smallest element of $\mathcal{C}$ that contains $H$. Also, by the definition of $\mathbf{K}_{\mathcal{C}}(H)$, it follows immediately that $H \subseteq \mathbf{K}_{\mathcal{C}}(H)$ for every $H \in \mathcal{P}(S)$.

Suppose that $H_1, H_2 \in \mathcal{P}(S)$ are such that $H_1 \subseteq H_2$. Since

$$\{L \in \mathcal{C} \mid H_2 \subseteq L\} \subseteq \{L \in \mathcal{C} \mid H_1 \subseteq L\},$$

we have

$$\bigcap\{L \in \mathcal{C} \mid H_1 \subseteq L\} \subseteq \bigcap\{L \in \mathcal{C} \mid H_2 \subseteq L\},$$

so $\mathbf{K}_{\mathcal{C}}(H_1) \subseteq \mathbf{K}_{\mathcal{C}}(H_2)$.

We have $\mathbf{K}_{\mathcal{C}}(H) \in \mathcal{C}$ for every $H \in \mathcal{P}(S)$ because $\mathcal{C}$ is a closure system. Therefore, $\mathbf{K}_{\mathcal{C}}(H) \in \{L \in \mathcal{C} \mid \mathbf{K}_{\mathcal{C}}(H) \subseteq L\}$, so $\mathbf{K}_{\mathcal{C}}(\mathbf{K}_{\mathcal{C}}(H)) \subseteq \mathbf{K}_{\mathcal{C}}(H)$. Since the reverse inclusion clearly holds, we obtain $\mathbf{K}_{\mathcal{C}}(\mathbf{K}_{\mathcal{C}}(H)) = \mathbf{K}_{\mathcal{C}}(H)$. □

**Definition 4.41.** *Let $\mathcal{C}$ be a closure system on a set $S$ and let $T$ be a subset of $S$. The* $\mathcal{C}$-set generated by *$T$ is the set $\boldsymbol{K}_{\mathcal{C}}(T)$.*

Note that $\mathbf{K}_{\mathcal{C}}(T)$ is the least set in $\mathcal{C}$ that includes $T$.

**Theorem 4.42.** *Let $S$ be a set. For every closure system $\mathcal{C}$ on $S$, we have $\mathcal{C} = \mathcal{C}_{\boldsymbol{K}_{\mathcal{C}}}$. For every closure operator $\boldsymbol{K}$ on $S$, we have $\boldsymbol{K} = \boldsymbol{K}_{\mathcal{C}_{\boldsymbol{K}}}$.*

*Proof.* Let $\mathcal{C}$ be a closure system on $S$ and let $H \subseteq M$. Then, we have the following equivalent statements:

1. $H \in \mathcal{C}_{\mathbf{K}_{\mathcal{C}}}$.
2. $\mathbf{K}_{\mathcal{C}}(H) = H$.
3. $H \in \mathcal{C}$.

The equivalence between (2) and (3) follows from the fact that $\mathbf{K}_{\mathcal{C}}(H)$ is the smallest element of $\mathcal{C}$ that contains $H$.

Conversely, let $\mathbf{K}$ be a closure operator on $S$. To prove the equality of $\mathbf{K}$ and $\mathbf{K}_{\mathcal{C}_{\mathbf{K}}}$, consider the following list of equal sets, where $H \subseteq S$:

1. $\mathbf{K}_{\mathcal{C}_{\mathbf{K}}}(H)$.
2. $\bigcap\{L \in \mathcal{C}_{\mathbf{K}} \mid H \subseteq L\}$.
3. $\bigcap\{L \in \mathcal{P}(S) \mid H \subseteq L = \mathbf{K}(L)\}$.
4. $\mathbf{K}(H)$.

We need to justify only the equality of the last two members of the list. Since $H \subseteq \mathbf{K}(H) = \mathbf{K}(\mathbf{K}(H))$, we have $\mathbf{K}(H) \in \{L \in \mathcal{P}(S) \mid H \subseteq L = \mathbf{K}(L)\}$. Thus, $\bigcap\{L \in \mathcal{P}(S) \mid H \subseteq L = \mathbf{K}(L)\} \subseteq \mathbf{K}(H)$. To prove the reverse inclusion, note that for every $L \in \{L \in \mathcal{P}(S) \mid H \subseteq L = \mathbf{K}(L)\}$, we have $H \subseteq L$, so $\mathbf{K}(H) \subseteq \mathbf{K}(L) = L$. Therefore, $\mathbf{K}(H) \subseteq \bigcap\{L \in \mathcal{P}(S) \mid H \subseteq L = \mathbf{K}(L)\}$. □

Theorem 4.42 shows the existence of a natural bijection between the set of closure operators on a set $S$ and the set of closure systems on $S$.

**Definition 4.43.** *Let $\mathcal{C}$ be a closure system on a set $S$ and let $T$ be a subset of $S$. The $\mathcal{C}$-closure of the set $T$ is the set $\boldsymbol{K}_{\mathcal{C}}(T)$.*

As we observed before, $\mathbf{K}_{\mathcal{C}}(T)$ is the smallest element of $\mathcal{C}$ that contains $T$.

*Example 4.44.* Let $\mathbf{K}$ be the closure operator given in Example 4.38. Since the closure system $\mathcal{C}_{\mathbf{K}}$ equals the range of $\mathbf{K}$, it follows that the members of $\mathcal{C}_{\mathbf{K}}$, the $\mathbf{K}$-closed sets, are $\emptyset$, $\mathbb{R}$, and all closed intervals $[a, b]$ with $a \leq b$. Thus, $\mathcal{C}_{\mathbf{K}}$ is the closure system $\mathcal{C}$ introduced in Example 4.33. Therefore, $\mathbf{K}$ and $\mathcal{C}$ correspond to each other under the bijection of Theorem 4.42.

For a relation $\rho$, on $S$ define $\rho^+$ as $\mathbf{K}_{TRAN(S)}(\rho)$. The relation $\rho^+$ is called the *transitive closure* of $\rho$ and is the least transitive relation containing $\rho$.

**Theorem 4.45.** *Let $\rho$ be a relation on a set $S$. We have*

$$\rho^+ = \bigcup\{\rho^n \mid n \in \mathbb{N} \text{ and } n \geq 1\}.$$

*Proof.* Let $\tau$ be the relation $\bigcup\{\rho^n \mid n \in \mathbb{N} \text{ and } n \geq 1\}$. We claim that $\tau$ is transitive. Indeed, let $(x, z), (z, y) \in \tau$. There exist $p, q \in \mathbb{N}$, $p, q \geq 1$ such that $(x, z) \in \rho^p$ and $(z, y) \in \rho^q$. Therefore, $(x, y) \in \rho^p \rho^q = \rho^{p+q} \subseteq \rho^+$, which shows that $\rho^+$ is transitive. The definition of $\rho^+$ implies that if $\sigma$ is a transitive relation such that $\rho \subseteq \sigma$, then $\rho^+ \subseteq \sigma$. Therefore, $\rho^+ \subseteq \tau$.

Conversely, since $\rho \subseteq \rho^+$ we have $\rho^n \subseteq (\rho^+)^n$ for every $n \in \mathbb{N}$. The transitivity of $\rho^+$ implies that $(\rho^+)^n \subseteq \rho^+$, which implies $\rho^n \subseteq \rho^+$ for every

$n \geq 1$. Consequently, $\tau = \bigcup\{\rho^n \mid n \in \mathbb{N} \text{ and } n \geq 1\} \subseteq \rho^+$. This proves the equality of the theorem. □

It is easy to see that the set of all reflexive and transitive relations on a set $S$, $REFTRAN(S)$, is also a closure system on the set of relations on $S$.

For a relation $\rho$ on $S$, define $\rho^*$ as $\mathbf{K}_{REFTRAN(S)}(\rho)$. The relation $\rho^*$ is called the *transitive-reflexive closure* of $\rho$ and is *the least transitive and reflexive relation* containing $\rho$. We have the following analog of Theorem 4.45.

**Theorem 4.46.** *Let $\rho$ be a relation on a set $S$. We have*

$$\rho^* = \bigcup\{\rho^n \mid n \in \mathbb{N}\}.$$

*Proof.* The argument is very similar to the proof of Theorem 4.45; we leave it to the reader. □

**Definition 4.47.** *Let $S$ be a set and let $F$ be a set of operations on $S$. A subset $P$ of $S$ is* closed under $F$, *or* $F$-closed, *if $P$ is closed under $f$ for every $f \in F$; that is, for every operation $f \in F$, if $f$ is $n$-ary and $p_0, \ldots, p_{n-1} \in P$, then $f(p_0, \ldots, p_{n-1}) \in P$.*

Note that $S$ itself is closed under $F$. Further, if $\mathcal{C}$ is a nonempty collection of $F$-closed subsets of $S$, then $\bigcap \mathcal{C}$ is also $F$-closed.

*Example 4.48.* Let $F$ be a set of operations on a set $S$. The collection of all $F$-closed subsets of a set $S$ is a closure system.

Let $S$ be a set. The dual of the poset $(\mathcal{P}(S), \subseteq)$ is the poset $(\mathcal{P}(S), \supseteq)$. Thus, the dual of the notion of closure operator on a set $S$, introduced in Definition 4.37, is a mapping $\mathbf{I} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ that satisfies the conditions

1. $U \supseteq \mathbf{I}(U)$ *(contraction)*,
2. $U \supseteq V$ implies $\mathbf{I}(U) \supseteq \mathbf{I}(V)$ *(monotonicity)*, and
3. $\mathbf{I}(\mathbf{I}(U)) = \mathbf{I}(U)$ *(idempotency)*,

for $U, V \in \mathcal{P}(S)$. Such a mapping is known as an *interior operator on the set $S$*.

The dual notion for the notion of closure system on a set introduced in Definition 4.32 is introduced next.

**Definition 4.49.** *An* interior system on a set $S$ *is a collection $\mathcal{I}$ of subsets of $S$ such that*
(i) $\emptyset \in \mathcal{I}$ *and,*
(ii) *for every subcollection $\mathcal{D}$ of $\mathcal{I}$ we have $\bigcup \mathcal{D} \in \mathcal{I}$.*

**Theorem 4.50.** *Let $\boldsymbol{I} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ be an interior operator. Define the family of sets $\mathcal{I}_{\boldsymbol{I}} = \{U \in \mathcal{P}(S) \mid U = \boldsymbol{I}(U)\}$. Then, $\mathcal{I}_{\boldsymbol{I}}$ is an interior system on $S$.*

*Conversely, if $\mathcal{I}$ is an interior system on the set $S$, define the mapping $\boldsymbol{I}_{\mathcal{I}} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ by $\boldsymbol{I}_{\mathcal{I}}(U) = \bigcup\{V \in \mathcal{I} \mid V \subseteq U\}$. Then, $\boldsymbol{I}_{\mathcal{I}}$ is an interior operator on the set $S$.*

*Moreover, for every interior system $\mathcal{I}$ on $S$, we have $\mathcal{I} = \mathcal{I}_{\boldsymbol{I}_{\mathcal{I}}}$. For every interior operator $\boldsymbol{I}$ on $S$, we have $\boldsymbol{I} = \boldsymbol{I}_{\mathcal{I}_{\boldsymbol{I}}}$.*

*Proof.* This statement follows by duality from Lemmas 4.39 and 4.40 and from Theorem 4.42. □

We refer to the sets in $\mathcal{I}_{\mathbf{I}}$ as the ***I*-open subsets of $S$**.

**Theorem 4.51.** *Let $\boldsymbol{K} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ be a closure operator on the set $S$. Then, the mapping $\boldsymbol{L} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ given by $\boldsymbol{L}(U) = S - \boldsymbol{K}(S - U)$ for $U \in \mathcal{P}(S)$ is an interior operator on $S$.*

*Proof.* Since $S - U \subseteq \mathbf{K}(S - U)$, it follows that $\mathbf{L}(U) \subseteq S - (S - U) = U$, which proves property (i) of Definition 4.49.

Suppose that $U \subseteq V$, where $U, V \in \mathcal{P}(S)$. Then, we have $S - V \subseteq S - U$, so $\mathbf{K}(S-V) \subseteq \mathbf{K}(S-U)$ by the monotonicity of closure operators. Therefore,

$$\mathbf{L}(U) = S - \mathbf{K}(S - U) \subseteq S - \mathbf{K}(S - V) = \mathbf{L}(V),$$

which proves the monotonicity of $\mathbf{L}$.

Finally, observe that we have $\mathbf{L}(\mathbf{L}(U)) \subseteq \mathbf{L}(U)$ because of the contraction property already proven for $\mathbf{L}$. Thus, we need only show that $\mathbf{L}(U) \subseteq \mathbf{L}(\mathbf{L}(U))$ to prove the idempotency of $\mathbf{L}$. This inclusion follows immediately from

$$\mathbf{L}(\mathbf{L}(U)) = \mathbf{L}(S - \mathbf{K}(S - U)) \supseteq \mathbf{L}(S - (S - U)) = \mathbf{L}(U).$$

□

From Theorem 4.51, by duality, we can prove that if $\mathbf{L}$ is an interior operator on a set $S$, then $\mathbf{K} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ defined as $\mathbf{K}(U) = S - \mathbf{L}(S - U)$ for $U \in \mathcal{P}(S)$ is a closure operator on the same set.

In Chapter 6, we extensively use closure and interior operators.

## 4.6 The Poset of Partitions of a Set

In Definition 1.113, we introduced the relation $\leq$ on the set of partitions $PART(S)$ of $S$. It is easy to verify that this is a partial order relation on $PART(S)$. Thus, the pair $(PART(S), \leq)$ is a poset. In this section, we study a few properties of this poset.

*Example 4.52.* The Hasse diagram of $(PART(\{1, 2, 3, 4\}), \leq)$ is given in Figure 4.3. To simplify this figure, we represent each nonempty subset of $\{1, 2, 3, 4\}$ as an increasing set of its elements and omit the outer braces; for instance, instead of $\{1, 2, 3\}$, we write 123.

**Theorem 4.53.** *Let $\pi, \sigma \in PART(S)$ such that $\pi \leq \sigma$. The partition $\sigma$* covers *the partition $\pi$ if and only if there exists a block $C$ of $\sigma$ that is the union of two blocks $B$ and $B'$ of $\pi$ and every block of $\sigma$ that is distinct of $C$ is a block of $\pi$.*

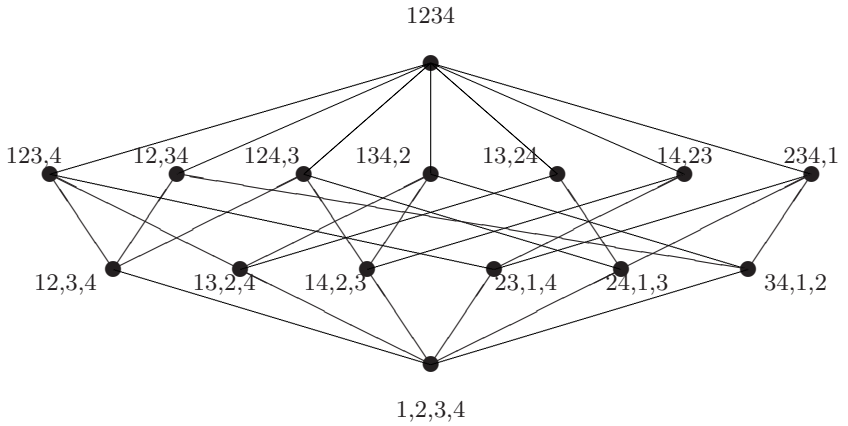



**Fig. 4.3.** The Hasse diagram of $(PART(\{1,2,3,4\}), \leq)$.

*Proof.* Suppose that $\sigma$ is a partition that covers the partition $\pi$. Since $\pi \leq \sigma$, every block of $\sigma$ is a union of blocks of $\pi$. Suppose that there exists a block $E$ of $\sigma$ that is the union of more than two blocks of $\pi$; that is, $E = \bigcup\{B_i \mid i \in I\}$, where $|I| \geq 3$, and let $B_{i_1}, B_{i_2}, B_{i_3}$ be three blocks of $\pi$ included in $E$. Consider the partitions

$$\sigma_1 = \{C \in \sigma \mid C \neq E\} \cup \{B_{i_1}, B_{i_2}, B_{i_3}\},$$
$$\sigma_2 = \{C \in \sigma \mid C \neq E\} \cup \{B_{i_1} \cup B_{i_2}, B_{i_3}\}.$$

It is easy to see that $\pi \leq \sigma_1 < \sigma_2 < \sigma$, which contradicts the fact that $\sigma$ covers $\pi$. Thus, each block of $\sigma$ is the union of at most two blocks of $\pi$.

Suppose that $\sigma$ contains two blocks $C'$ and $C''$ that are unions of two blocks of $\pi$, namely $C' = B_{i_0} \cup B_{i_1}$ and $C'' = B_{i_2} \cup B_{i_3}$. Define the partitions

$$\sigma' = \{C \in \sigma \mid C \notin \{C', C''\}\} \cup \{C', B_{i_2}, B_{i_3}\},$$
$$\sigma'' = \{C \in \sigma \mid C \notin \{C', C''\}\} \cup \{B_{i_1}, B_{i_2}, C''\}.$$

Since $\pi < \sigma', \sigma'' < \sigma$, this contradicts the fact that $\sigma$ covers $\pi$. Thus, we obtain the conclusion of the theorem. □

The poset $(PART(S), \leq)$ has $\alpha_S$ as its first element and $\omega_S$ as its largest.

We shall prove that for two partitions $\pi$ and $\sigma$ the infimum and the supremum of the set $\{\pi, \sigma\}$ always exist. To facilitate the description of the partitions $\inf\{\pi, \sigma\}$ and $\sup\{\pi, \sigma\}$, we introduce the graph of a pair of partitions.

**Definition 4.54.** *Let $\pi, \sigma \in PART(S)$, where $\pi = \{B_i \mid i \in I\}$ and $\sigma = \{C_j \mid j \in J\}$. The* graph of the pair $(\pi, \sigma)$ *is the graph*

$$\mathcal{G}_{\pi,\sigma} = (\{B_i \mid i \in I\} \cup \{C_j \mid j \in J\}, E),$$

*where $E$ consists of those two-element sets $\{B_i, C_j\}$ such that $B_i \cap C_j \neq \emptyset$.*

*Example 4.55.* Let $S = \{a_i \mid 1 \leq i \leq 12\}$ and let $\pi = \{B_i \mid 1 \leq i \leq 5\}$ and $\sigma = \{C_j \mid 1 \leq j \leq 4\}$, where

$$\begin{array}{ll} B_1 = \{a_1, a_2\}, & C_1 = \{a_2, a_4\}, \\ B_2 = \{a_3, a_4, a_5\}, & C_2 = \{a_1, a_3, a_5, a_6, a_7\}, \\ B_3 = \{a_6, a_7\}, & C_3 = \{a_8, a_{11}\}, \\ B_4 = \{a_8, a_9, a_{10}\}, & C_4 = \{a_9, a_{10}, a_{12}\}, \\ B_5 = \{a_{11}, a_{12}\}. & \end{array}$$

The graph $G_{\pi,\sigma}$ is shown in Figure 4.4.

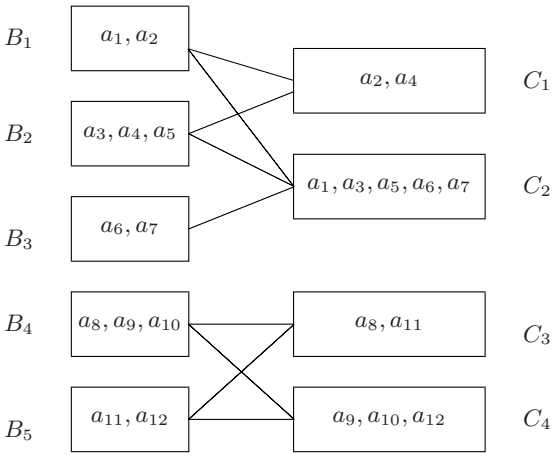



**Fig. 4.4.** The graph $G_{\pi,\sigma}$.

**Theorem 4.56.** *Let $\pi, \sigma \in PART(S)$, where $\pi = \{B_i \mid i \in I\}$ and $\sigma = \{C_j \mid j \in J\}$. The partition $\inf\{\pi, \sigma\}$ exists and is given by*

$$\inf\{\pi, \sigma\} = \{B_i \cap C_j \mid i \in I, j \in J \text{ and } B_i \cap C_j \neq \emptyset\}.$$

*Proof.* It is clear that the collection of sets

$$\tau = \{B_i \cap C_j \mid i \in I, j \in J \text{ and } B_i \cap C_j \neq \emptyset\}$$

is a partition of $S$ and that $\tau \leq \pi$ and $\tau \leq \sigma$.

Let $\tau'$ be a partition of $S$ such that $\tau' \leq \pi$ and $\tau' \leq \sigma$ and let $D$ be a block of $\tau'$. There are $B_i \in \pi$ and $C_j \in \sigma$ such that $D \subseteq B_i$ and $D \subseteq C_j$, so $D \subseteq B_i \cap C_j$. Therefore, $\tau' \leq \tau$ and this proves that $\tau' = \inf\{\pi, \sigma\}$. □

The partition $\inf\{\pi, \sigma\}$ will be denoted by $\pi \wedge \sigma$. Note that the blocks of $\pi \wedge \sigma$ correspond to the edges of the graph $\mathcal{G}_{\pi,\sigma}$.

*Example 4.57.* If $\pi$ and $\sigma$ are the partitions introduced in Example 4.55, then the partition $\pi \wedge \sigma$ consists of nine blocks that correspond to the edges of the graph:

$$\begin{array}{lll} B_1 \cap C_1 = \{a_2\}, & B_1 \cap C_2 = \{a_1\}, & B_2 \cap C_1 = \{a_4\}, \\ B_2 \cap C_2 = \{a_3, a_5\}, & B_3 \cap C_2 = \{a_6, a_7\}, & B_4 \cap C_3 = \{a_8\}, \\ B_4 \cap C_4 = \{a_9, a_{10}\}, & B_5 \cap C_3 = \{a_{11}\}, & B_5 \cap C_4 = \{a_{12}\}. \end{array}$$

**Theorem 4.58.** *Let $\pi$ and $\sigma \in PART(S)$, where $\pi = \{B_i \mid i \in I\}$ and $\sigma = \{C_j \mid j \in J\}$. The $\sup\{\pi, \sigma\}$ exists in the poset $(PART(S), \leq)$, and the blocks of the partition $\sup\{\pi, \sigma\}$ are the unions of the blocks that belong to connected components of the graph $G_{\pi,\sigma}$.*

*Proof.* The connected components of the graph $\mathcal{G}_{\pi,\sigma}$ form a partition of the set of vertices of the graph. Let $\tau$ be the partition of $S$ whose blocks are the unions of the blocks that belong to connected components of the graph $G_{\pi,\sigma}$.

Let $D$ be a block of $\tau$ and let $\{B_{i_1}, \ldots, B_{i_p}\}$ and $\{C_{j_1}, \ldots, C_{j_q}\}$ be the sets of blocks of $\pi$ and $\sigma$, respectively, that are included in $\tau$. We claim that

$$\bigcup\{B_{i_k} \mid 1 \leq k \leq p\} = \bigcup\{C_{j_h} \mid 1 \leq h \leq q\}.$$

Indeed, let $x \in \bigcup\{B_{i_k} \mid 1 \leq k \leq p\}$. There exists a block $B_{i_\ell}$ such that $x \in B_{i_\ell}$. Also, there is a block $C$ of $\sigma$ such that $x \in C$. Since $x \in B_{i_\ell} \cap C$, it follows that there exists an edge $(B_{i_\ell}, C)$ in $\mathcal{G}_{\pi,\sigma}$, so $C$ belongs to the same connected component as $B_{i_\ell}$; that is, $C = C_{j_g}$ for some $g$, $1 \leq g \leq q$. Therefore,

$$\bigcup\{B_{i_k} \mid 1 \leq k \leq p\} \subseteq \bigcup\{C_{j_h} \mid 1 \leq h \leq q\}.$$

The reverse inclusion can be shown in a similar manner. This proves the needed equality, which can now be written

$$D = \bigcup\{B_{i_k} \mid 1 \leq k \leq p\} = \bigcup\{C_{j_h} \mid 1 \leq h \leq q\}.$$

It is clear that we have both $\pi \leq \tau$ and $\sigma \leq \tau$.

Suppose now that $\tau'$ is a partition such that $\pi \leq \tau'$ and $\pi \leq \tau'$. Let $B \in \pi$ and $C \in \sigma$ be two blocks that have a nonempty intersection. If $x \in B \cap C$, then both $B$ and $C$ are included in the block of $\tau'$ that contains $x$. In other words, if in $\mathcal{G}_{\pi,\sigma}$ an edge exists that joins $B$ and $C$, then they are both included in the same block of $\tau'$. This property can be extended to paths: if there is a path in $\mathcal{G}_{\pi,\sigma}$ that joins a block $B$ of $\pi$ to a block $C$ of $\sigma$, then the union of all $\pi$-blocks and of all the $\sigma$-blocks along this path is included in a block $E$ of $\tau'$. The argument, by induction on the length of the path, is immediate and is omitted. Thus, every block of $\tau$, which is a union of all $\pi$-blocks that

belong to a connected component (and of all $\sigma$-blocks that belong to the same connected component), is included in a block of $\tau'$. Therefore, $\tau \leq \tau'$, and this proves that $\tau = \sup\{\pi, \sigma\}$. □

Let $\pi, \sigma \in PART(S)$ and let $\tau = \sup\{\pi, \sigma\}$. We have $(x, y) \in \rho_\tau$ if and only if $\{x, y\}$ is enclosed in the same connected component of the graph $\mathcal{G}_{\pi,\sigma}$; that is, if and only if there exists an alternating sequence of blocks of $\pi$ and $\sigma$ – $B_{i_1}, C_{j_1}, B_{i_2}, C_{j_2}, \ldots, B_{i_r}, C_{j_s}$ – such that $x \in B_{i_1}$ and $y \in C_{j_s}$. This is equivalent to the existence of a sequence of elements $z_0, z_1, \ldots, z_m$ of $S$ such that $z_0 = x$, $z_m = y$, and $(z_i, z_{i+1}) \in \rho_\pi$ or $(z_i, z_{i+1}) \in \rho_\sigma$ for every $i$, $0 \leq i \leq m-1$.

The partition $\sup\{\pi, \sigma\}$ will be denoted by $\pi \vee \sigma$.

*Example 4.59.* The graph of the partitions $\pi, \sigma$ introduced in Example 4.55 has two connected components that correspond to the blocks,

$$\begin{aligned} D_1 &= \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\} \\ &= B_1 \cup B_2 \cup B_3 \\ &= C_1 \cup C_2, \\ D_2 &= \{a_8, a_9, a_{10}, a_{11}, a_{12}\} \\ &= B_4 \cup B_5 \\ &= C_3 \cup C_4, \end{aligned}$$

of the partition $\pi \vee \sigma$.

**Theorem 4.60.** *Let $S$ be a set and let $\rho, \rho' \in EQS(S)$. We have $\pi_\rho \wedge \pi_{\rho'} = \pi_{\rho \cap \rho'}$.*

*Proof.* Indeed, note that $\rho \cap \rho'$ is an equivalence on $S$, and the equivalence classes of this equivalence (that is, the blocks of the partition $\pi_{\rho \cap \rho'}$) are the nonempty intersections of the blocks of $\rho$ and $\rho'$. The definition of the infimum of two partitions shows that the set of blocks of $\pi_\rho \wedge \pi_{\rho'}$ is exactly the same, which gives the equality of the theorem. □

## 4.7 Chains and Antichains

The main notions of this section are introduced next.

**Definition 4.61.** *Let $(S, \leq)$ be a poset. A* chain *of $(S, \leq)$ is a subset $T$ of $S$ such that for every $x, y \in T$ such that $x \neq y$ we have either $x < y$ or $y < x$. If the set $S$ is a chain, we say that $(S, \leq)$ is a* totally ordered set *and the relation $\leq$ is a* total order.

*If $\boldsymbol{s} \in \mathbf{Seq}(S)$ (or $\boldsymbol{s} \in \mathbf{Seq}_\infty(S)$) and for every $i, j \in \mathbb{N}$ we have $\boldsymbol{s}(i) < \boldsymbol{s}(j)$ or $\boldsymbol{s}(j) < \boldsymbol{s}(i)$, we refer to the sequence $\boldsymbol{s}$ as a chain in $S$; if $\boldsymbol{s}(i) \leq \boldsymbol{s}(j)$ or $\boldsymbol{s}(j) \leq \boldsymbol{s}(i)$ for every $i, j \in \mathbb{N}$, then we say that $\boldsymbol{s}$ is a* multichain *in $(S, \leq)$.*

If $S = \{x_1, \ldots, x_n\}$, the total order whose diagram is given in Figure 4.5 is denoted by $TO(x_1, \ldots, x_n)$.

**Fig. 4.5.** Hasse diagram of a total order on $S = \{x_1, x_2, \ldots, x_n\}$.

Let $(S, \leq)$ be a poset. The elements $x, y$ of $S$ are *incomparable* if we have neither $x \leq y$ nor $y \leq x$. This is denoted by $x \parallel y$. It is easy to see that "$\parallel$" is a symmetric and irreflexive relation. The set of pairs of incomparable elements of a poset $(S, \leq)$ is

$$INC(S, \leq) = \{(x, y) \in S \times S \mid x \not\leq y \text{ and } y \not\leq x\}.$$

**Definition 4.62.** *An* antichain *of $(S, \leq)$ is a subset $U$ of $S$ such that, for every two distinct elements $x, y \in U$, we have $x \parallel y$.*

*Example 4.63.* The set of real numbers equipped with the usual partial order $(\mathbb{R}, \leq)$ is a chain since, for every $x, y \in \mathbb{R}$, we have either $x \leq y$ or $y \leq x$.

*Example 4.64.* In the poset $(\mathbb{N}, \delta)$, the set of all prime numbers is an antichain since if $p$ and $q$ are two distinct primes, we have neither $(p, q) \in \delta$ nor $(q, p) \in \delta$.

*Example 4.65.* If $S$ is a finite set such that $|S| = n$, the set of subsets of $S$ that contain $k$ elements (for a fixed $k$, $k \leq |S|$) is an antichain in the poset $(\mathcal{P}(S), \subseteq)$ that contains $\binom{n}{k}$ elements.

*Example 4.66.* If $(S, \leq)$ is a poset, then both $MIN(S, \leq)$ and $MAX(S, \leq)$ are maximal antichains of $(S, \leq)$ (with respect to set inclusion).

Every finite chain of a poset has a least element and a greatest element. Indeed, by Theorem 4.25, a finite chain has a minimal element and a maximal element. Since the notions of minimal and maximal elements in a chain

coincide with the notions of least element and largest element, respectively, it statement follows.

Not every poset is a chain, as shown in the next example.

*Example 4.67.* The poset $(\mathcal{P}(S), \subseteq)$ considered in Example 4.12 is not a chain; elements of $\mathcal{P}(S)$ such as $\{a, b\}$ and $\{b, c\}$ are incomparable.

The poset from Example 4.13 is not a chain since it contains incomparable elements (for instance, $4 \parallel 6$). However, the subset $\{2, 4, 8\}$ is a chain, as can be easily seen. Thus, a poset $(S, \leq)$ that is not a chain itself may very well contain subsets that are chains with respect to the trace of the partial order of the set itself.

Denote by $CHAINS(S)$ the set of chains of a poset $(S, \leq)$. We use the poset $(CHAINS(S), \subseteq)$, where the partial order relation is the set inclusion.

**Theorem 4.68.** *If $\{U_i \mid i \in I\}$ is a chain of the poset $(CHAINS(S), \subseteq)$ (that is, a chain of chains of $(S, \leq)$), then $\bigcup\{U_i \mid i \in I\}$ is itself a chain of $(S, \leq)$ (that is, a member of $(CHAINS(S), \subseteq)$).*

*Proof.* Let $x, y \in \bigcup\{U_i \mid i \in I\}$. There are $i, j \in I$ such that $x \in U_i$ and $y \in U_j$ and we have either $U_i \subseteq U_j$ or $U_j \subseteq U_i$. In the first case, we have either $x_i \leq x_j$ or $x_j \leq x_i$ because both $x$ and $y$ belong to the chain $U_j$. The same conclusion can be reached in the second case when both $x$ and $y$ belong to the chain $U_i$. So, in any case, $x$ and $y$ are comparable, which proves that $\bigcup\{U_i \mid i \in I\}$ is a chain of $(S, \leq)$. □

**Definition 4.69.** *A* well-ordered *poset is a poset for which every nonempty subset has a least element.*

A well-ordered set is necessarily a chain. Indeed, consider the well-ordered set $(S, \leq)$ and $x, y \in S$. Since the set $\{x, y\}$ must have a least element, we have either $x \leq y$ or $y \leq x$.

*Example 4.70.* The set of natural numbers is well-ordered. This property of natural numbers is known as the *well-ordering principle.*

> **(Well-Ordering Axiom)** Given any set $S$, there is a binary relation $\rho$ such that $(S, \rho)$ is a well-ordered set.

The set $(\mathbb{R}, \leq)$ is not well-ordered, despite the fact that it is a chain, since it contains subsets such as $(0, 1) = \{x \mid x \in R, 0 < x < 1\}$ that do not have a least element.

**Definition 4.71.** *Let "$<$" be the strict partial order of the poset $(S, \leq)$. An* infinite descending sequence *in a poset $(S, \rho)$ is an infinite sequence $\boldsymbol{s} \in \mathbf{Seq}_\infty(S)$ such that $\boldsymbol{s}(n+1) < \boldsymbol{s}(n)$ for all $n \in \mathbb{N}$.*

*An* infinite ascending sequence *in a poset $(S, \rho)$ is an infinite sequence $\boldsymbol{s} \in \mathbf{Seq}_\infty(S)$ such that $\boldsymbol{s}(n) < \boldsymbol{s}(n+1)$ for all $n \in \mathbb{N}$.*

*A poset with no infinite descending sequences is called* Artinian. *A poset with no infinite ascending sequences is called* Noetherian.

Clearly, the range of every infinite ascending or descending sequence is a chain.

*Example 4.72.* The poset $(\mathbb{N}, \delta)$ is Artinian. Indeed, suppose that $\mathbf{s}$ is an infinite descending sequence of natural numbers. If $\mathbf{s}(0) \neq 0$, then the natural number $\mathbf{s}(0)$ has an infinite set of divisors $\{\mathbf{s}(0), \mathbf{s}(1), \ldots\}$. If $\mathbf{s}(0) = 0$, in view of the fact that any natural number is a divisor of 0, we obtain the impossibility of an infinite descending sequence by applying the same argument to $\mathbf{s}(1)$. However, this poset is not Noetherian. For instance, the sequence $\mathbf{z} : \mathbb{N} \longrightarrow \mathbb{N}$ defined by $\mathbf{z}(n) = 2^n$ for $n \in \mathbb{N}$ is an infinite ascending sequence.

A generalization of well-ordered posets is considered in the next definition.

**Definition 4.73.** *A* well-founded poset *is a partially ordered set where every nonempty subset has a minimal element.*

Since the least element of a subset is also a minimal element, it is clear that a well-ordered set is also well-founded. However, the inverse is not true; for instance, not every finite set is well-ordered.

**Theorem 4.74.** *A poset $(S, \rho)$ is well-founded if and only if it is Artinian.*

*Proof.* Let $(S, \rho)$ be a well-founded poset, and suppose that $\mathbf{s}$ is an infinite descending sequence in this poset. The set $T = \{\mathbf{s}(n) \mid n \in \mathbb{N}\}$ has no minimal element since, for every $\mathbf{s}(k) \in T$, we have $(\mathbf{s}(k+1), \mathbf{s}(k)) \in \rho_1$, which contradicts the well-foundedness of $(S, \rho)$.

Conversely, assume that $(S, \rho)$ is Artinian; that is, there is no infinite descending sequence in $(S, \rho)$. Suppose that $K$ is a nonempty subset of $S$ without minimal elements. Let $x_0$ be an arbitrary element of $K$. Such an element exists since $K$ is not empty. Since $x_0$ is not minimal, there is $x_1 \in K$ such that $(x_1, x_0) \in \rho$. Since $x_1$ is not minimal, there is $x_2 \in K$ such that $(x_2, x_1) \in \rho$, etc., and this construction can continue indefinitely. In this way, we can build an infinite descending sequence $\mathbf{s} : \mathbb{N} \longrightarrow S$, where $\mathbf{s}(n) = x_n$ for $n \in \mathbb{N}$. □

Theorem 4.74 implies immediately that any finite poset is well-founded.

*Example 4.75.* We will show that the poset $(\mathbb{N} \times \mathbb{N}, \preceq)$ is well-founded.

If $(m, n_0) \succ (m, n_1) \succ \ldots$ is a descending chain of pairs having the same first component, then $n_0 > n_1 > \ldots$ is a descending chain of natural numbers and such a chain is finite. Therefore, $(m, n_0) \succ (m, n_1) \succ \ldots$ must be a finite chain.

Consider now an arbitrary descending chain,

$$(p_0, q_0) \succ (p_1, q_1) \succ \ldots,$$

in $(\mathbb{N} \times \mathbb{N}, \preceq)$. We have $p_0 \geq p_1 \geq \ldots$, and in this sequence we may have only finite "constant" fragments $p_k = p_{k+1} = \cdots = p_{k+l}$. Therefore, the chain of the first components of the pairs of the sequence $(p_0, q_0) \succ (p_1, q_1) \succ \ldots$ is ultimately decreasing, and this shows that the chain is finite. Thus, this poset is Artinian and therefore, by Theorem 4.74, it is well-founded.

**Definition 4.76.** *Let $(S, \leq)$ be a poset that has a least element denoted by 0.*
*The* height of an element $x \in S$ *(denoted by* $\mathsf{height}(x)$*) is the least upper bound of the lengths of the chains of the form $0 < x_1 < \cdots < x_k = x$.*

If $x$ is an atom of a poset that has the least element 0, then $\mathsf{height}(x) = 1$.

**Definition 4.77.** *A poset $(S, \leq)$ satisfies the* Jordan-Dedekind *condition if all maximal chains between the same elements have the same finite length.*

*Example 4.78.* The poset $(M_5, \leq)$ whose Hasse diagram is shown in Figure 4.6(a) satisfies the Jordan-Dedekind condition; the poset $(N_5, \leq)$ shown in Figure 4.6(b) fails this condition because it contains two maximal chains $0 < x < y < 1$ and $0 < z < 1$ of different lengths between 0 and 1.

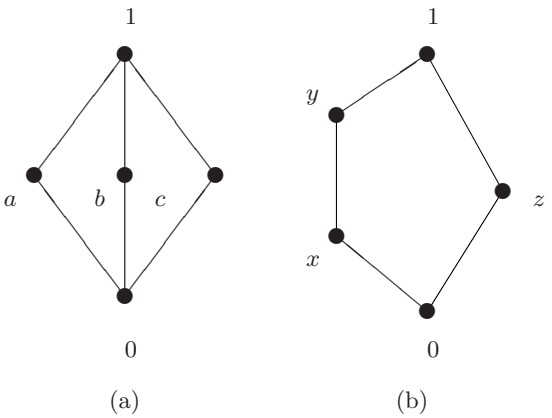



**Fig. 4.6.** Hasse diagrams of posets $(M_5, \leq)$ and $(N_5, \leq)$.

**Theorem 4.79.** *Let $(S, \leq)$ be a poset that has finite chains and has the least element 0. $(S, \leq)$ satisfies the Jordan-Dedekind condition if and only if the following conditions are satisfied:*
(i) *$x < y$ implies $\mathsf{height}(x) < \mathsf{height}(x)$ and*
(ii) *$y$ covers $x$ implies $\mathsf{height}(y) = \mathsf{height}(x) + 1$*
*for every $x, y \in S$.*

*Proof.* If the height function satisfies the conditions of the theorem, then any chain between the elements $x$ and $y$ has length $\mathsf{height}(y) - \mathsf{height}(x)$, so the Jordan-Dedekind condition is satisfied. Conversely, if the Jordan-Dedekind condition holds, then $\mathsf{height}(x)$ is the length of any maximal chain between 0 and $x$ and the conditions of the theorem follow immediately. □

Theorem 4.79 suggests the following definition.

**Definition 4.80.** *A* graded poset *is a triple $(S, \leq, h)$, where $(S, \leq)$ is a poset and $h : S \longrightarrow \mathbb{N}$ is a function that satisfies the conditions*
(i) *$x < y$ implies $h(x) < h(x)$ and*
(ii) *$y$ covers $x$ implies $h(y) = h(x) + 1$,*
*for every $x, y \in S$. The function $h$ is referred to as the* grading function.

*The set $L_k = \{x \in S \mid h(x) = k\}$ is called the $k$-th* level of the poset *$(S, \leq, h)$.*

*Example 4.81.* Define the function $h : M_5 \longrightarrow \mathbb{N}$ by $h(0) = 0$, $h(a) = h(b) = h(c) = 1$, and $h(1) = 2$. The triple $(M_5, \leq, h)$ is a graded poset. Its levels are

$$L_0 = \{0\},$$
$$L_1 = \{a, b, c\},$$
$$L_2 = \{1\}.$$

**Definition 4.82.** *Let $(S, \leq)$ be a finite poset. The* height *of $(S, \leq)$, denoted by $\mathsf{height}(S, \leq)$, is the maximal number of elements of a chain. The* width *of $(S, \leq)$, width$(S, \leq)$, is the maximal number of elements of an antichain. The* length *of $(S, \leq)$ is the number $\mathsf{length}(S, \leq) = \mathsf{height}(S, \leq) - 1$.*

*Example 4.83.* Let $S = \{s_1, \ldots, s_n\}$ be a finite set such that $|S| = n$. The poset $(\mathcal{P}(S), \subseteq)$ has height $n + 1$ since a maximal chain has the form $(\emptyset, \{s_{i_1}\}, \{s_{i_1}, s_{i_2}\}, \ldots, S)$, where $(s_{i_1}, s_{i_2}, \ldots, s_{i_n})$ is a permutation of $S$. Its width is $\binom{n}{\lfloor n/2 \rfloor}$.

It is clear that if a finite poset $(S, \leq)$ contains an antichain $U$ such that $|U| = m$, then $S$ is the union of at least $m$ chains since no two elements of an antichain may belong to the same chain.

**Theorem 4.84 (Dilworth's Theorem).** *If $(S, \leq)$ is a finite nonempty poset such that width$(S, \leq) = m$, then there is a partition of $S$ into $m$ chains.*

*Proof.* The argument is by strong induction on $n = |S|$. If $n = 1$, then the statement holds trivially.

Suppose that the statement holds for sets with fewer than $n$ elements, and let $(S, \leq)$ be a poset with $|S| = n$.

Let $C$ be a maximal chain in $(S, \leq)$. Two cases may occur:

(i) If no antichain of $(S - C, \leq)$ has $m$ elements, then, by the induction hypothesis, there exists a partition of $S - C$ into $m - 1$ chains, so there is a partition of $S$ into $m$ chains.
(ii) If $S - C$ has an antichain $U = \{u_1, \ldots, u_m\}$, define the sets $\mathsf{UP}_U$ and $\mathsf{DOWN}_U$ as

$$\mathsf{UP}_U = \{x \in S | x \geq u_i \text{ for some } u_i \in U\},$$
$$\mathsf{DOWN}_U = \{x \in S | x \leq u_i \text{ for some } u_i \in U\}.$$

Note that $S = \mathsf{UP}_U \cup \mathsf{DOWN}_U$ since otherwise $S$ would contain an antichain with more than $m$ elements. Since $(S, \leq)$ is a finite poset, the chain $C$ has a largest element $t_1$ and a smallest element $t_0$. We have the strict inclusions $\mathsf{UP}_U \subset S$ and $\mathsf{DOWN}_U \subset S$ because $t_1 \notin \mathsf{DOWN}_U$ and $t_0 \notin \mathsf{UP}_U$. Thus, both $\mathsf{DOWN}_U$ and $\mathsf{UP}_U$ have fewer than $n$ elements.

By the induction hypothesis, we can decompose both $\mathsf{UP}_U$ and $\mathsf{DOWN}_U$ as partitions of chains, $UP_U = \bigcup_{i=1}^{m} C_{\geq}^i$ and $\mathsf{DOWN}_U = \bigcup_{i=1}^{m} C_{\leq}^i$, where $u_i \in C_{\geq}^i \cap C_{\leq}^i$. Note that $u_i$ is the least element of $C_{\geq}^i$ and the greatest element of $C_{\leq}^i$. Therefore, $C_{\geq}^i \cup C_{\leq}^i$ is a chain, which gives the desired result. □

Next, we state a related statement using antichains.

**Theorem 4.85.** *If $(S, \leq)$ is a finite nonempty poset such that $\mathsf{height}(S, \leq) = m$, then there is a partition of $S$ into $m$ antichains.*

*Proof.* We construct a sequence of finite posets $(S_i, \leq_i)$ for $0 \leq i \leq k-1$. The first poset is $(S_0, \leq_0) = (S, \leq)$.

Suppose that we defined the nonempty poset $(S_i, \leq_i)$. Consider the antichain $U_{i+1} = MAX(S_i, \leq_i)$ and the poset $(S_{i+1}, \leq_{i+1})$, where $S_{i+1} = S_i - U_{i+1}$ and $\leq_{i+1} = (\leq_i)_{S_{i+1}}$. The process halts when $S_k = S_{k-1} - U_k = \emptyset$. It is clear that the $U_1, \ldots, U_k$ are $k$ pairwise disjoint antichains in $(S, \leq)$ and that $S = \bigcup_{i=1}^{k} U_i$.

Since no two members of an antichain may belong to the same chain and $S$ contains a chain having $m$ elements, it follows that any partition of $S$ into antichains requires at least $m$ antichains. Therefore, we have $m \leq k$, which means that we need to show only that $k \leq m$.

To prove that $k \leq m$, we construct a chain $x_1 < x_2 < \cdots < x_k$ in the poset $(S, \leq)$ beginning with $x_k$. Choose $x_k$ to be an arbitrary element of $U_k$. If $x_j \in U_j$ for $i \leq j \leq k$, then choose $x_{i-1} \in U_{i-1}$ such that $x_i < x_{i-1}$. This choice is possible because otherwise $x_i \in U_{i-1} = MAX(S_{i-1}, \leq_{i-1})$, which is contradictory because $x_i \in U_i$. This proves that $\{x_1, \ldots, x_k\}$ is a chain, so $\mathsf{height}(S, \leq) = m \geq k$. □

## 4.8 Poset Product

Let $I$ be a set, and $(S, \rho)$ be a poset. The partial order $\rho$ generates a partial order $\rho$ on the set of functions $I \longrightarrow S$ using the following definition. If $f, g : I \longrightarrow S$, we have $(f, g) \in \rho$ if $(f(i), g(i)) \in \rho$ for every $i \in I$.

The relation $\rho$ on $I \longrightarrow S$ is a partial order. We verify only the antisymmetry and leave for the reader the proofs of the reflexivity and transitivity. Assume that $(f, g), (g, f) \in \rho$ for $f, g : I \longrightarrow M$. We have $(f(i), g(i)) \in \rho$ and $(g(i), f(i)) \in \rho$ for every $i \in I$. Therefore, taking into account the antisymmetry of $\rho$, we obtain $f(i) = g(i)$ for all $i \in I$; hence, $f = g$, which proves the antisymmetry of $\rho$.

For a set of functions $F \subseteq I \longrightarrow S$, define the subset $F(i)$ of $S$ as $S(i) = \{f(i) \mid f \in F\}$ for $i \in I$.

**Theorem 4.86.** *The subset $F$ of the poset $(I \longrightarrow S, \rho)$ has a supremum if and only if $\sup F(i)$ exists for every $i \in I$ in the poset $(S, \rho)$.*

*Proof.* Suppose that $\sup F(i)$ exists for every $i \in I$ in the poset $(S, \rho)$. Define the mapping $g : I \longrightarrow S$ by $g(i) = \sup F(i)$ for every $i \in I$. We claim that $g$ is $\sup F$.

If $f \in F$, then $(f(i), g(i)) \in \rho$ for every $i \in I$ because of the definition of $g$. This shows that $(f, g) \in \rho$; hence, $g$ is an upper bound of $F$. Let $h$ be an upper bound of $F$. For every $f \in F$, we have $(f(i), h(i)) \in \rho$ for $i \in I$. The definition of $g$ implies $(g(i), h(i)) \in \rho$ for $i \in I$; hence, $g = \sup F$.

Conversely, assume that $k = \sup F$ exists in the poset $(I \longrightarrow S, \rho)$. We prove that $k(i)$ is $\sup F(i)$ for every $i \in I$ in the poset $(S, \rho)$.

The definition of $k$ implies that, for every $f \in F$, we have $(f, k) \in \rho$; that is, $(f(i), k(i)) \in \rho$ for every $i \in I$. Therefore, $k(i)$ is an upper bound of the set $F(i)$ for every $i \in I$.

Let $l_i$ be an upper bound for $F(i)$ for $i \in I$. Define the function $l : I \longrightarrow S$ as $l(i) = l_i$ for $i \in I$. Clearly, $l$ is an upper bound of the set $F$ in the poset $(I \longrightarrow S, \rho)$, and therefore $(k, l) \in \rho$. This, in turn, means that $(k(i), l(i)) = (k(i), l_i) \in \rho$, which shows that $\sup F(i)$ exists and is equal to $k(i)$. □

**Definition 4.87.** *The* product of the posets $\{(S_i, \leq_i) \mid i \in I\}$ *is the poset $(D, \leq)$, where $D = \prod_{i \in I} S_i$ and "$\leq$" is the partial order introduced above on $D$. When $I = \{1, \ldots, n\}$, the product will be denoted by*

$$(S_1, \leq_1) \times \cdots \times (S_n, \leq_n)$$

*or by $\prod_{i \in I}(S_i, \leq_i)$.*

**Theorem 4.88.** *Let $\{(S_i, \leq_i) \mid i \in I\}$ be a family of partially ordered sets. If $H \subseteq \prod_{i \in I} S_i$, then in the product poset, $\sup H$ ($\inf H$) exists if and only if $\sup p_i(H)$ ($\inf p_i(H)$, respectively) exists for every $i \in I$. Moreover, if $y = \sup H$ ($y = \inf H$), then $p_i(y) = \sup p_i(H)$ ($p_i(y) = \inf p_i(H)$) for every $i \in I$.*

*Proof.* Assume that $y_i = \sup p_i(H)$ exists for every $i \in I$. We need to prove that the element $y$ of $\prod_{i\in I} S_i$ defined by $p_i(y) = y_i$ is $\sup H$.

Consider an arbitrary element $z \in H$. Since $p_i(z) \in p_i(H)$, we have $p_i(z) \leq_i y_i$, that is, $p_i(z) \leq_i p_i(y)$ for every $i \in I$. This means that $z \leq y$, which shows that $y$ is an upper bound of $H$.

Suppose now that $v$ is an arbitrary upper bound of $H$. To show that $y$ is $\sup H$, we need to prove that $y$ is the least upper bound of $H$; that is, $y \leq v$ or, equivalently, $p_i(y) \leq_i p_i(v)$ for every $i \in I$.

If $v$ is an upper bound of $H$, then $p_i(v)$ is an upper bound of $p_i(H)$. Since $p_i(y) = y_i = \sup p_i(H)$, we obtain immediately $p_i(y) \leq_i p_i(v)$ for every $i \in I$.

Conversely, suppose that $\sup H$ exists. Let $y = \sup H$ and let $y_i = p_i(y)$ for every $i \in I$. We have $x_i \in p_i(H)$ if there is $x \in H$ such that $p_i(x) = x_i$. Since $x \leq y$, it follows that $x_i \leq_i p_i(y)$, which shows that $p_i(y)$ is an upper bound for $p_i(H)$.

Let $w_i$ be an arbitrary upper bound of $p_i(H)$ for every $i \in I$. There is $w \in \prod_{i\in I} S_i$ such that $p_i(w) = w_i$, and we have $y \leq w$ because $w$ is an upper bound for $H$. Consequently, $p_i(y) \leq_i p_i(w)$, and this means that $y_i = \sup p_i(H)$ for every $i \in I$.

The statement for inf follows by dualization. □

Another kind of partial order that can be introduced on $S_1 \times \cdot \times S_n$ is defined next.

**Theorem 4.89.** *For $f, g \in S_1 \times \cdots \times S_n$, define $f \preceq g$ if $f = g$ or if there is $k$, $1 \leq k \leq n$, such that $f(k) \neq g(k)$, $f(i) = g(i)$ for $1 \leq i < k$ and $f(k) <_k g(k)$.*

*The relation $\preceq$ is a partial order on $S_1 \times \cdots \times S_n$.*

*Proof.* The relation $\preceq$ is obviously reflexive. Suppose now that $f \preceq g$ and $g \preceq f$ and that $f \neq g$. There are $k, h \in \mathbb{N}$ such that $f(i) = g(i)$ for $1 \leq i < k$, $f(k) <_k g(k)$, and $f(i) = g(i)$ for $1 \leq i < h$, $f(h) <_h g(h)$. If $k < h$, this leads to a contradiction since we cannot have $f(k) <_k g(k)$ and $f(k) = g(k)$. The case $h < k$ also results in a contradiction. For $k = h$, the previous supposition implies $f(k) <_k g(k)$ and $g(k) <_k f(k)$, which is contradictory because "$<_k$" is a strict partial order.

Assume that $f \preceq g$ and $g \preceq l$ and that $f \neq g$, $g \neq l$. There are $k, h \in \mathbb{N}$ such that $f(i) = g(i)$ for $1 \leq i < k$, $f(k) <_k g(k)$, and $g(i) = l(i)$ for $1 \leq i < h$, $g(h) <_h l(h)$. Define $p$ as being the least of the numbers $k, h$. For $1 \leq i < p$, we have $f(i) = g(i) = l(i)$. In addition, we have $f(p) \leq_p l(p)$. Three cases may occur:

1. $f(p) = g(p)$ and $g(p) <_p l(p)$ (when $k > h$),
2. $f(p) <_p g(p)$ and $g(p) = l(p)$ (when $k < h$), and
3. $f(p) <_p g(p))$ and $g(p) <_p l(p))$ (when $k = h$).

If $f = l$, then we have $f \preceq l$. Therefore, we can assume that $f \neq l$. In the first two cases mentioned above, this would imply immediately $f \preceq l$ because of the fact that $f(p) <_p l(p)$. The same conclusion can be reached in the third case because of the transitivity of the strict partial order $<_p$. □

We refer to the partial order "$\preceq$" as the *lexicographic partial order* on $S_1 \times \cdots \times S_n$.

Let $\{(S_i, \leq_i) \mid 1 \leq i \leq n\}$ be a family of totally ordered posets. The product poset $\prod_{i=1}^{n}(S_i, \leq_i)$ is not necessarily a total order; however, the lexicographic product $(S_1 \times \cdots \times S_n, \preceq)$ is a total order (see Exercise 25).

*Example 4.90.* Consider the totally ordered set $(\{0,1\}, \leq)$, whose Hasse diagram is given in Figure 4.7(a). The Hasse diagram of the poset $(S \times S, \leq)$ is shown in Figure 4.7(b).

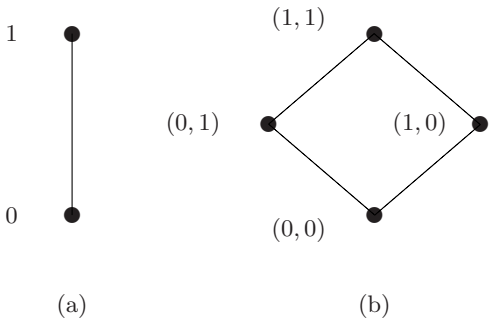



**Fig. 4.7.** Hasse diagrams of $(\{0,1\}, \leq)$ and $(\{0,1\}^2, \leq)$.

On the other hand, the Hasse diagram of the poset $(\{0,1\}^2, \preceq)$ given in Figure 4.8 shows that "$\preceq$" is a total order on $\{0,1\}^2$.

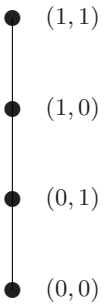



**Fig. 4.8.** Hasse diagram of $(\{0,1\}^2, \preceq)$.

If $S_1 = \cdots S_n = S$, then we obtain the poset $(\mathbf{Seq}_n(S), \preceq)$.

## 4.9 Functions and Posets

Let $(S, \leq)$ and $(T, \leq)$ be two posets.

**Definition 4.91.** *A* morphism *between* $(S, \leq)$ *and* $(T, \leq)$ *or a* monotonic mapping *between* $(S, \leq)$ *and* $(T, \leq)$ *is a mapping* $f : S \longrightarrow T$ *such that* $u, v \in S$ *and* $u \leq v$ *imply* $f(u) \leq f(v)$.

*A mapping* $g : S \longrightarrow T$ *is* antimonotonic *if* $u, v \in S$ *and* $u \leq v$ *imply* $g(u) \geq g(v)$.

*The mapping* $f$ *is* strictly monotonic *if* $u < v$ *implies* $f(u) < f(v)$, *where "*$<$*" is the strict partial order associated with the partial order "*$\leq$*".*

Note that $g : S \longrightarrow T$ is antimonotonic if and only if $g$ is a monotonic mapping between the poset $(S, \leq)$ and the dual $(T, \geq)$ of the poset $(T, \leq)$.

*Example 4.92.* Consider a set $M$, the poset $(\mathcal{P}(M), \subseteq)$, and the functions $f, g : (\mathcal{P}(M))^2 \longrightarrow \mathcal{P}$, defined by $f(K, H) = K \cup H$ and $g(K, H) = K \cap H$, for $K, H \in \mathcal{P}(M)$. If the Cartesian product is equipped with the product partial order, then both $f$ and $g$ are monotonic. Indeed, if $(K_1, H_1) \subseteq (K_2, H_2)$, we have $K_1 \subseteq K_2$ and $H_1 \subseteq H_2$, which implies that

$$f(K_1, H_1) = K_1 \cup H_1 \subseteq K_2 \cup H_2 = f(K_2, H_2).$$

The argument for $g$ is similar, and it is left to the reader.

*Example 4.93.* Let $\{(S_i, \rho_i) \mid i \in I\}$ be a collection of posets and let

$$\left(\prod_{i \in I} S_i, \rho\right)$$

be the product of these posets. The projections $p_i : \prod_{i \in I} S_i \longrightarrow S_i$ are monotonic mappings, as the reader will easily verify.

*Example 4.94.* Let $(M, \rho)$ be an arbitrary poset. Any function $f : S \longrightarrow M$ is monotonic when considered between the posets $(S, \iota_S)$ and $(M, \rho)$.

**Theorem 4.95.** *Let* $(P, \leq), (R, \leq), (S, \leq)$ *be three posets and let* $f : P \longrightarrow R$, $g : R \longrightarrow S$ *be two monotonic mappings. The mapping* $gf : P \longrightarrow S$ *is also monotonic.*

*Proof.* Let $x, y \in P$ be such that $x \leq y$. In view of the monotonicity of $f$, we have $f(x) \leq f(y)$, and this implies $(g(f(x)) \leq g(f(y))$ because of the monotonicity of $g$. Therefore, $gf$ is monotonic. □

Let $(P, \leq)$ and $(R, \leq)$ be two posets. For a monotonic function $f : P \longrightarrow R$, the quotient set, $P/\mathbf{ker}(f)$ can also be organized as a poset. Indeed, if $[x], [y] \in P/\mathbf{ker}(f)$, then we define $[x] \leq [y]$ if $f(x) \leq f(y)$. This partial order o $P/\mathbf{ker}(f)$ is well-defined because if $x' \in [x]$ and $y' \in [y]$, we have $(f(x'), f(y')) = (f(x), f(y))$.

**Theorem 4.96.** *The mapping $g : P \longrightarrow P/\boldsymbol{ker}(f)$ defined by $g(x) = [x]$ for $x \in P$ is a monotonic mapping between the posets $(P, \leq)$ and $(P/\boldsymbol{ker}(f), \leq)$.*

*Proof.* The argument is straightforward, and it is left to the reader as an exercise. □

Let $f : S \longrightarrow T$ be a monotonic bijection between the posets $(S, \leq)$ and $(T, \leq)$. As we have seen in Chapter 1, the inverse $f^{-1}$ is also a bijection. Nevertheless, the inverse *is not* necessarily monotonic, as follows from the next example.

*Example 4.97.* Let $(M_5, \leq)$ and $(N_5, \leq)$ be the posets whose Hasse diagrams are given in Figure 4.6, and consider the mapping $f : M_5 \longrightarrow N_5$ defined by $f(0) = 0$, $f(a) = y$, $f(b) = x$, $f(c) = z$, and $f(1) = 1$. The inverse bijection $f^{-1}$ is not monotonic because we have $x \leq y$ in $(N_5, \leq)$ and $(f^{-1}(x), f^{-1}(y)) = (b, a)$ and $b \not\leq a$ in $(M_5, \leq)$.

Let $(R, \leq)$ and $(S, \leq)$ be two posets. The previous considerations justify the following definition.

**Definition 4.98.** *A* poset isomorphism *between the posets $(R, \leq)$ and $(S, \leq)$ is a monotonic bijective mapping $f : R \longrightarrow S$ for which the inverse mapping $f^{-1}$ is also monotonic.*

*If a poset isomorphism exists between the posets $(P, \leq)$ and $(S, \leq)$, then we refer to these posets as* isomorphic.

*Example 4.99.* Let $\{p_1, p_2, \ldots, p_n\}$ be the first $n$ primes, $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, etc. Let $m = p_1 \cdots p_n$ be their product and let $D_m$ be the set of all divisors of $m$. Consider an arbitrary set $A = \{a_1, \ldots, a_n\}$ having $n$ elements.

The posets $(\mathcal{P}(A), \subseteq)$ and $(D_m, \delta)$ are isomorphic. Indeed, define the mapping $f : \mathcal{P}(A) \longrightarrow D_m$ by $f(\emptyset) = 1$ and $f(\{a_{i_1}, \ldots, a_{i_k}\}) = p_{i_1} \cdots p_{i_k}$.

The mapping $f$ is bijective. Indeed, for any divisor $h$ of $m$, we have $h = p_{i_1} \cdots p_{i_k}$ and therefore $h = f(\{a_{i_1}, \ldots, a_{i_k}\})$, which shows that $f$ is surjective.

If $f(\{a_{i_1}, \ldots, a_{i_k}\}) = f(\{a_{j_1}, \ldots, a_{j_l}\})$, then $p_{i_1} \cdots p_{i_k} = p_{j_1} \cdots p_{j_l}$. This gives $k = l$ and $i_1 = j_1, \ldots, i_k = j_k$; hence, $\{a_{i_1}, \ldots, a_{i_k}\} = \{a_{j_1}, \ldots, a_{j_l}\}$, which proves that $f$ is injective.

The mapping $f$ is monotonic because if $\{a_{i_1}, \ldots, a_{i_k}\} \subseteq \{a_{j_1}, \ldots, a_{j_l}\}$,

$$\{i_1, \ldots, i_k\} \subseteq \{j_1, \ldots, j_l\},$$

and this means that the number $p_{i_1} \cdots p_{i_k}$ divides $p_{j_1} \cdots p_{j_l}$.

The inverse mapping $g : D_m \longrightarrow \mathcal{P}(A)$ is also monotonic; we leave the argument to the reader.

Monotonic functions map chains to chains, as we show next.

**Theorem 4.100.** *Let $(P, \leq)$ and $(R, \leq)$ be two posets and $f : P \longrightarrow R$ be a monotonic function. If $L \subseteq P$ is a chain in $(P, \leq)$, then $f(L)$ is a chain in $(R, \leq)$.*

*Proof.* Let $u, v \in f(L)$ be two elements of $f(L)$. There exist $x, y \in L$ such that $f(x) = u$ and $f(y) = v$. Since $L$ is a chain, we have either $x \leq y$ or $y \leq x$. In the former case, the monotonicity of $f$ implies $u \leq v$; in the latter situation, we have $v \leq u$. □

## 4.10 Posets and the Axiom of Choice

A statement equivalent to the Axiom of Choice, known as Zorn's lemma, can be stated in the framework of posets.

**Zorn's Lemma:** If every chain of a poset $(S, \leq)$ has an upper bound, then $S$ has a maximal element.

**Theorem 4.101.** *The following three statements are equivalent for a poset* $(S, \leq)$*:*

(i) *If every chain of* $(S, \leq)$ *has an upper bound, then* $S$ *has a maximal element (Zorn's Lemma).*

(ii) *If every chain of* $(S, \leq)$ *has a least upper bound, then* $S$ *has a maximal element.*

(iii) $S$ *contains a chain that is maximal with respect to set inclusion (Hausdorff maximality principle).*

*Proof.* (i) implies (ii) is immediate.

(ii) implies (iii): Let $(CHAINS(S), \subseteq)$ be the poset of chains of $S$ ordered by set inclusion. By Theorem 4.68, every chain $\{U_i \mid i \in I\}$ of the poset $(CHAINS(S), \subseteq)$ has a least upper bound $\bigcup\{U_i \mid i \in I\}$ in the poset $(CHAINS(S), \subseteq)$. Therefore, by (ii), $(CHAINS(S), \subseteq)$ has a maximal element that is a chain of $(S, \leq)$ that is maximal with respect to set inclusion.

(iii) implies (i): Suppose that $S$ contains a chain $W$ that is maximal with respect to set inclusion and that every chain of $(S, \leq)$ has an upper bound. Let $w$ be an upper bound of $W$.

If $w \in W$, then $w$ is a maximal element of $S$. Indeed, if this were not the case, then $S$ would contain an element $t$ such that $w < t$ and $W \cup \{t\}$ would be a chain that would strictly include $W$.

If $w \notin W$, then $W \cup \{w\}$ would be a chain strictly including $W$, which, again, would contradict the maximality of $W$. Thus, $w$ is a maximal element of $(S, \leq)$. □

Denote by $PORD(S)$ the collection of partial order relations on the set $S$.

**Definition 4.102.** *Let* $\rho, \rho' \in PORD(S)$*. The partial order* $\rho'$ *is an* extension *of* $\rho$ *if* $(x, y) \in \rho$ *implies* $(x, y) \in \rho'$*. Equivalently, we shall say that* $\rho'$ extends $\rho$*.*

An important consequence of Zorn's lemma is the next statement, which shows that any partial order defined on a set can be extended to a total order on the same set.

**Theorem 4.103 (Szpilrajn's Theorem).** *Let $(S, \leq)$ be a poset. There is a total order $\leq'$ on $S$ that is an extension of $\leq$.*

*Proof.* Let $PORD(S, \leq)$ be the set of partial order relations that can be defined on the set $S$ and contain the relation "$\leq$"; clearly, the relation "$\leq$" itself is a member of $PORD(S, \leq)$. We will apply Zorn's lemma to the poset $(PORD(S, \leq), \subseteq)$.

Let $\mathcal{R} = \{\rho_i \mid i \in I\}$ be a chain of $(PORD(S, \leq), \subseteq)$; that is, a chain of partial orders $\rho_i$ relative to set inclusion such that $x \leq y$ implies $(x, y) \in \rho_i$ for every $i \in I$ and all $x, y \in S$. We claim that the relation $\rho = \bigcup \mathcal{R}$ is a partial order on $S$.

Indeed, since $\iota_S \subseteq \leq \subseteq \rho_i$ for $i \in I$ we have $\iota_S \subseteq \rho$, so $\rho$ is a reflexive relation. To prove that $\rho$ is antisymmetric let $x, y \in S$ be two elements such that $(x, y) \in \rho$ and $(y, x) \in \rho$. By the definition of $\rho$, there exist $i, j \in I$ such that $(x, y) \in \rho_i$ and $(y, x) \in \rho_j$. Since $\mathcal{R}$ is a chain, we have either $\rho_i \subseteq \rho_j$ or $\rho_j \subseteq \rho_i$. In the first case, both $(x, y)$ and $(y, x)$ belong to $\rho_j$, so $x = y$ because of the antisymmetry of $\rho_j$; in the second case, the same conclusion follows because $(x, y)$ and $(y, x)$ belong to $\rho_i$. Thus, $\rho$ is indeed antisymmetric.

We leave it to the reader to prove the transitivity of $\rho$. Thus, $\rho$ is a partial order that includes "$\leq$", and the arbitrary chain $\mathcal{R}$ has an upper bound. By Zorn's lemma the poset $(PORD(S, \leq), \subseteq)$ has a maximal element $\leq'$. We now prove that $\leq'$ is a total order.

Suppose that $(u, v)$ and $(v, u)$ are two distinct ordered pairs of elements of $S$ such that $u \not\leq' v$ and $v \not\leq' u$. We show that this supposition leads to a contradiction.

Let $\leq_1$ be the relation on $S$ given by

$$\begin{aligned}\leq_1 = \{(x, y) \in S \times S \mid x \leq' y\} \cup \{(u, v)\} \\ \cup \{(z, v) \in S \times \{v\} \mid z \leq' v\} \cup \{(u, t) \in \{u\} \times S \mid u \leq' t\}.\end{aligned}$$

Since $\iota_S \subseteq \leq' \subseteq \leq_1$, it follows that $\leq_1$ is reflexive.

To prove the antisymmetry of $\leq_1$, suppose that $p \leq_1 q$ and $q \leq_1 p$. Since $v \not\leq' u$, it follows that $(p, q) \neq (u, v)$. Thus, the following cases may occur:

(i) If $p \leq' q$ and $q \leq' p$, then $p = q$ by the antisymmetry of $\leq'$.
(ii) If $p = u$, we have $u \leq_1 q$ and $q \leq_1 u$. By the definition of $\leq_1$, this implies $u \leq' q$ and $q \leq' u$, respectively, so $q = u = p$.
(iii) If $q = v$, we have $p \leq_1 v$ and $v \leq_1 p$, which imply $p \leq' v$ and $v \leq' p$, respectively. Thus, $p = v = q$.

We leave the proof of transitivity for "$\leq_1$" to the reader.

Note that $\leq'$ is strictly included in $\leq_1$ because $u \not\leq' v$. This contradicts the maximality of the partial order $\leq'$, so $\leq'$ must be a total order. □

*Example 4.104.* Consider the poset $(N_5, \leq)$ introduced in Example 4.78. The posets $(N_5, \leq_i)$, where $1 \leq i \leq 3$ whose Hasse diagrams are shown in Figure 4.9(a)–(c) are such that $\leq \subset \leq_i$ and $\leq_i$ is a total order for $1 \leq i \leq 3$. Also, it is easy to see that we have actually $\leq = \leq_1 \cap \leq_3$.

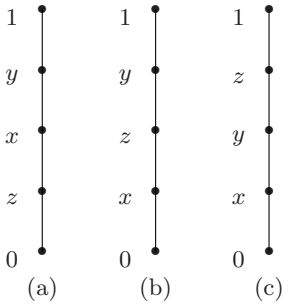



**Fig. 4.9.** Hasse diagrams of three total orders on the set $\{0, x, y, z, 1\}$.

**Corollary 4.105.** *Let $(S, \leq)$ be a poset and let $x$ and $y$ be two incomparable elements in $(S, \leq)$. There exists a total order $\leq'$ on $S$ that extends $\leq$ such that $x \leq' y$ and a total order $\leq''$ that extends $\leq$ such that $y \leq'' x$.*

*Proof.* This statement follows immediately from Szpilrajn's theorem. □

## 4.11 Locally Finite Posets and Möbius Functions

**Definition 4.106.** *Let $(S, \leq)$ be a poset and let $x, y \in S$ be such that $x \leq y$. The* closed interval *of $(S, \leq)$ defined by $x, y$ is the set*

$$[x, y] = \{t \in S \mid x \leq t \leq y\}.$$

*In addition, we define the* open interval $(x, y)$ *as*

$$(x, y) = \{t \in S \mid x < t < y\}$$

*and the* semiclosed *(or* semiopen*) intervals $[x, y)$ and $(x, y]$ by*

$$[x, y) = \{t \in S \mid x \leq t < y\},$$
$$(x, y] = \{t \in S \mid x < t \leq y\},$$

*respectively.*

Note that if $x = y$, then $[x, x] = \{x\}$, while $(x, x) = \emptyset$.

**Definition 4.107.** *A poset $(S, \leq)$ is* locally finite *if every closed interval of $(S, \leq)$ is finite.*

*Example 4.108.* The poset $(\mathbb{N}, \leq)$ is locally finite. Indeed, if $[p, q]$ is a closed interval of this poset, then $[p, q]$ is a finite set that consists of $q - p + 1$ natural numbers.

*Example 4.109.* The poset $(\mathbb{N}, \delta)$ introduced in Example 1.29 is locally finite. Indeed, if $p$ divides $q$, then $[p, q]$ is a finite set that contain all multiples of $p$ that divide $q$. For example, the closed interval $[2, 12]$ contains the numbers $2, 4, 6$ and $12$.

Let $(S, \leq)$ be a locally finite poset and let $\mathcal{A}(S, \leq)$ be the set of all functions of the form $f : S \times S \longrightarrow \mathbb{R}$ such that $x \not\leq y$ implies $f(x, y) = 0$ for $x, y \in S$. We refer to $\mathcal{A}(S, \leq)$ as the *incidence algebra* of the poset $(S, \leq)$.

Note that if $f \in \mathcal{A}(S, \leq)$ and $x > y$ or $x \parallel y$, then $f(x, y) = 0$.

**Definition 4.110.** *Let $(S, \leq)$ be a locally finite poset and let $f, g \in \mathcal{A}(S, \leq)$ be two functions. Their* convolution product *is the function $h : S \times S \longrightarrow \mathbb{R}$ defined by*

$$h(x, y) = \begin{cases} \sum_{z \in [x,y]} f(x, z)g(z, y) & \text{if } x \leq y, \\ 0 & \text{otherwise,} \end{cases}$$

*for $x, y \in S$. The function $h$ will be denoted by $f * g$.*

**Lemma 4.111.** *The operation $*$ is well-defined on the set $\mathcal{A}(S, \leq)$; further, "$*$" is associative on $\mathcal{A}(S, \leq)$ and has the* Kronecker function $\mathsf{k}$ *defined by*

$$\mathsf{k}(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise,} \end{cases}$$

*for $x, y \in S$ as its unit element.*

*Proof.* Suppose that $h = f * g$, where $f, g \in \mathcal{A}(S, \leq)$. If $x \not\leq y$, then $h(x, y) = 0$, so $h \in \mathcal{A}(S, \leq)$.

Let $e, f, g$ be three functions of $\mathcal{A}(S, \leq)$. We claim that $(e*f)*g = e*(f*g)$. Suppose that $x \leq z$. Then, we have

$$\begin{aligned}
((e * f) * g)(x, z) &= \sum_{y \in [x,z]} (e * f)(x, y)g(y, z) \\
&= \sum_{y \in [x,z]} \left( \sum_{u \in [x,y]} e(x, u)f(u, y) \right) g(y, z) \\
&= \sum_{y \in [x,z]} \sum_{u \in [x,y]} e(x, u)f(u, y)g(y, z) \\
&= \sum_{y \in [x,z]} \sum_{u \in [x,z]} e(x, u)f(u, y)g(y, z) \\
&\quad \text{(because if } u > y \text{ we have } f(u, y) = 0) \\
&= \sum_{u \in [x,z]} \sum_{y \in [x,z]} e(x, u)f(u, y)g(y, z).
\end{aligned}$$

On the other hand, we can write

$$\begin{aligned}(e*(f*g))(x,z) &= \sum_{u\in[x,z]} e(x,u)(f*g)(u,z)\\ &= \sum_{u\in[x,z]} e(x,u)\sum_{y\in[u,z]} f(u,y)g(y,z)\\ &= \sum_{u\in[x,z]} e(x,u)\sum_{y\in[x,z]} f(u,y)g(y,z),\\ &\qquad \text{(because if } u>y \text{ we have } f(u,y)=0)\end{aligned}$$

for $x, z \in S$, which shows that $*$ is associative.

If $f \in \mathcal{A}(S,\leq)$ and $x \leq y$, then we can write

$$\begin{aligned}(f*\mathsf{k})(x,y) &= \sum_{z\in[x,y]} f(x,z)\mathsf{k}(z,y)\\ &= f(x,y)\end{aligned}$$

for $x, y \in S$. Thus, $f * \mathsf{k} = f$. A similar argument shows that $\mathsf{k} * f = f$. This allows us to conclude that $\mathsf{k}$ is indeed the unit with respect to the $*$ operation. □

Let $\mathcal{I}(S,\leq) = \{[x,y] \mid x, y \in S \text{ and } x \leq y\} \cup \{\emptyset\}$ be the set of intervals of the poset $(S,\leq)$ to which we add the empty set. A useful point of view (see [128]) is to regard the incidence algebra of $(S,\leq)$ as consisting of formal sums of the form $\sum\{f(x,y)\cdot[x,y] \mid [x,y] \in \mathcal{I}(S,\leq) - \{\emptyset\}\}$. Define the product of two intervals as

$$[x,y][u,v] = \begin{cases}[x,v] & \text{if } y = u,\\ \emptyset & \text{otherwise.}\end{cases}$$

Further, we assume that the product of formal sums is distributive with respect to addition of these sums. Let $f, g \in \mathcal{A}(S,\leq)$ be two functions and let $\hat{f}$ and $\hat{g}$ be their corresponding formal sums,

$$\begin{aligned}\hat{f} &= \sum\{f(x,y)\cdot[x,y] \mid [x,y] \in \mathcal{I}(S,\leq)\},\\ \hat{g} &= \sum\{g(u,v)\cdot[u,v] \mid [u,v] \in \mathcal{I}(S,\leq)\}.\end{aligned}$$

Then, it is immediate that

$$\hat{f}\hat{g}(x,z) = \sum_{x\leq y\leq z} f(x,y)g(y,z)[x,z],$$

so the usual product of the formal sums $\hat{f}\hat{g}$ corresponds to the convolution product of $f$ and $g$.

**Theorem 4.112.** *Let $(S, \leq)$ be a locally finite poset. A function $f \in \mathcal{A}(S, \leq)$ has an inverse relative to the operation $*$ if and only if $f(x, x) \neq 0$ for every $x \in S$.*

*Proof.* Suppose that there exists an inverse $f'$ of $f$ (that is, $f * f' = f' * f = \mathsf{k}$) which yields $(f * f')(x, x) = \mathsf{k}(x, x) = 1$ for every $x$. Since $(f * f')(x, x) = \sum_{z \in [x,x]} f(x, z)f'(z, x) = f(x, x)f'(x, x)$, it follows that $f(x, x) \neq 0$.

To prove the converse implication, we first show the existence of a left inverse of $f$; that is, a function $f' : S \times S \to \mathbb{R}$ such that $f' * f = \mathsf{k}$. For $x \leq y$, we must have $\sum_{z \in [x,y]} f'(x, z)f(z, y) = \mathsf{k}(x, y)$. This implies $f'(x, x)f(x, x) = 1$ and $\sum_{z \in [x,y]} f'(x, z)f(z, y) = 0$ if $x \neq y$. Thus, we must have

$$f'(x, x) = \frac{1}{f(x, x)}, \tag{4.3}$$

$$f'(x, y) = -\frac{1}{f(y, y)} \sum_{z \in [x,y)} f'(x, z)f(z, y), \tag{4.4}$$

when $x \leq y$ and

$$f'(x, y) = 0,$$

when $x \not\leq y$. Equalities (4.3) and (4.4) give an inductive definition of $f'$ because the poset $(S, \leq)$ is locally finite.

To verify that $f'$ is a left inverse of $f$, suppose that $x < y$. Then,

$$\begin{aligned}(f' * f)(x, y) &= \sum_{z \in [x,y]} f'(x, z)f(z, y)\\ &= \sum_{z \in [x,y)} f'(x, z)f(z, y) + f'(x, y)f(y, y) = 0.\end{aligned}$$

If $x = y$, then $(f' * f)(y, y) = 1$ and $x \not\leq y$ implies $(f' * f)(x, y) = 0$. Therefore, $f' * f = \mathsf{k}$.

The function $f'$ is also a right inverse of $f$. Let $h = f * f'$. We have shown above that every function of $\mathcal{A}(S, \leq)$ has a left inverse, so let $h'$ be the left inverse of $h$. Thus, we have $f * f' = h = \mathsf{k} * h = (h' * h) * h = h' * (f * f') * (f * f') = h' * f * \mathsf{k} * f' = h' * f * f' = h' * h = \mathsf{k}$, which proves that $f'$ is also a right inverse of $f$. Thus, $f'$ is the inverse of $f$. □

If the inverse of $f \in \mathcal{A}(S, \leq)$ exists, we will denote it by the common notation $f^{-1}$.

**Corollary 4.113.** *Let $(S, \leq)$ be a locally finite poset and let $\mathcal{IA}(S, \leq)$ be the set of invertible functions of $\mathcal{A}(S, \leq)$. Then $(\mathcal{IA}(S, \leq), \{\mathsf{k}, *, ^{-1}\})$ is a group.*

*Proof.* This is a mere restatement of Theorem 4.112. □

Let $(S, \leq)$ be a locally finite poset and let $\zeta : S \times S \longrightarrow \mathbb{R}$ be the *Riemann function* defined by

$$\zeta(x,y) = \begin{cases} 1 & \text{if } x \leq y, \\ 0 & \text{otherwise,} \end{cases}$$

for $x, y \in S$. Clearly, $\zeta \in \mathcal{A}(S, \leq)$, so the function $\zeta^{-1}$ exists by Corollary 4.113. This inverse, known as the *Möbius function*, is denoted by $\mu$ and its values can be computed from Equalities (4.3) and (4.4) as

$$\begin{aligned} \mu(x,x) &= \frac{1}{\zeta(x,x)} = 1, \\ \mu(x,y) &= -\sum_{z \in [x,y)} \mu(x,z)\zeta(z,y) \\ &= -\sum_{z \in [x,y)} \mu(x,z), \end{aligned}$$

for $x < y$; for $x \not\leq y$, we have $\mu(x,y) = 0$.

The special role played by $\mu$ is discussed next.

**Theorem 4.114 (Möbius Inversion Theorem).** *Let $(S, \leq)$ be a locally finite poset that has the least element 0. If $f, g : S \longrightarrow \mathbb{R}$ are two real-valued functions such that*

$$g(x) = \sum_{0 \leq z \leq x} f(z),$$

*then*

$$f(x) = \sum_{0 \leq z \leq x} g(z)\mu(z,x)$$

*for $x \in S$.*

*Proof.* Starting from the functions $f, g : S \longrightarrow \mathbb{R}$, define the functions $F, G \in \mathcal{A}(S, \leq)$ by

$$\begin{aligned} F(0,x) = f(x), G(0,x) = g(x) \\ F(u,x) = G(u,x) = 0, \text{ if } u > 0. \end{aligned}$$

The equality $g(x) = \sum_{0 \leq z \leq x} f(z)$ can be written as

$$G(0,x) = \sum_{0 \leq z \leq x} F(0,z)\zeta(z,x),$$

where $\zeta$ is Riemann's function. We also have $G(u,x) = \sum_{u \leq z \leq x} F(u,z)\zeta(z,x)$ for $u > 0$ because in this case $G(u,x) = 0$ and $F(u,z) = 0$. Thus, $G = F * \zeta$. Since $\mu$ is the inverse of $\zeta$ in $\mathfrak{IA}(S, \leq)$, it follows that $F = G * \mu$. Consequently,

$$\begin{aligned} f(x) &= F(0,x) \\ &= \sum_{0 \leq z \leq x} G(0,z)\mu(z,x) \\ &= \sum_{0 \leq z \leq x} g(z)\mu(z,x), \end{aligned}$$

which is the desired equality. □

Now let $(S, \leq)$ be a poset that has the greatest element 1. By applying the Möbius inversion theorem to its dual $(S, \leq^{-1}) = (S, \geq)$ we obtain the following dual form of the theorem.

**Theorem 4.115 (Möbius Dual Inversion Theorem).** *Let $(S, \leq)$ be a locally finite poset that has the greatest element 1. If $f, g : S \longrightarrow \mathbb{R}$ are two real-valued functions such that*

$$g(x) = \sum_{x \leq z \leq 1} f(z),$$

*then*

$$f(x) = \sum_{x \leq z \leq 1} g(z)\mu(z, x)$$

*for $x \in S$.*

*Proof.* This statement follows immediately from Theorem 4.114. □

*Example 4.116.* Let $M$ be a finite set and let $(\mathcal{P}(M), \subseteq)$ be the poset of all its subsets. The Möbius function of this poset is given by

$$\mu(A, B) = \begin{cases} (-1)^{|B|-|A|} & \text{if } A \subseteq B, \\ 0 & \text{otherwise,} \end{cases}$$

for $A, B \in \mathcal{P}(M)$.

Let $A, B \in \mathcal{P}(M)$ be such that $A \subseteq B$. We shall prove that $\mu(A, B) = (-1)^{|B|-|A|}$ by induction on $n = |B| - |A|$.

In the basis case $n = 0$, so $A = B$, which implies $\mu(A, B) = 1$, thus verifying the equality above. Suppose that the equality holds for sets that differ by fewer than $n$ elements and that $|B| - |A| = n$. Then, by the definition of the Möbius function, we have

$$\begin{aligned} \mu(A, B) &= -\sum_{C \in [A, B)} \mu(A, C), \\ &= -\sum_{C \in [A, B)} (-1)^{|C|-|A|}. \end{aligned}$$

Note that there are $2^n - 1$ sets $C$ in $[A, B)$. Namely, there are $\binom{n}{k}$ sets $C$ such that $|C| - |A| = k$. Therefore,

$$\sum_{C \in [A, B)} (-1)^{|C|-|A|} = \sum k = 0^{n-1}(-1)^k \binom{n}{k}.$$

Choosing $x = -1$ in the identity

$$(x+1)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$$

implies $0 = \sum_{k=0}^{n} \binom{n}{k}(-1)^k$, which yields the equality

$$\sum_{k=0}^{n-1} \binom{n}{k}(-1)^k = (-1)^{n+1}.$$

Thus, $\mu(A, B) = (-1)^{n+2} = (-1)^n = (-1)^{|B|-|A|}$.

## Exercises and Supplements

1. Define the relation $\leq$ on the set $\mathbb{N}^n$ by $(p_1, \ldots, p_n) \leq (q_1, \ldots, q_n)$ if $p_i \leq q_i$ for $1 \leq i \leq n$. Prove that $(\mathbb{N}^n, \leq)$ is a partially ordered set.
2. Let $S$ and $T$ be two sets and let $\sqsubseteq$ be the relation on $S \rightsquigarrow T$ defined by $f \sqsubseteq g$ if $\mathrm{Dom}(f) \subseteq \mathrm{Dom}(g)$ and $f(s) = g(s)$ for every $s \in \mathrm{Dom}(f)$. Prove that $\sqsubseteq$ is a partial order on $S \rightsquigarrow T$.
3. Prove that a binary relation $\rho$ on a set $S$ is a strict partial order on $S$ if and only if it is irreflexive, transitive, and antisymmetric.
4. Let $(S, \leq)$ be a poset. An *order ideal* is a subset $I$ of $S$ such that $x \in I$ and $y \leq x$ implies $y \in I$. If $\mathcal{I}(S, \leq)$ is the collection of order ideals of $(S, \leq)$, prove that $\mathcal{K} \subseteq \mathcal{I}(S, \leq)$ implies $\bigcap \mathcal{K} \in \mathcal{I}(S, \leq)$. Further, argue that $S \in \mathcal{I}(S, \leq)$.
5. Let $(S, \leq)$ be a poset. An *order filter* is a subset $F$ of $S$ such that $x \in F$ and $y \geq x$ implies $y \in F$. If $\mathcal{F}(S, \leq)$ is the collection of order filters of $(S, \leq)$, prove that $\mathcal{K} \subseteq \mathcal{F}(S, \leq)$ implies $\bigcap \mathcal{K} \in \mathcal{F}(S, \leq)$. Further, show that $S \in \mathcal{I}(S, \leq)$.
6. Let $(S, \leq)$ be a finite poset. Prove that $S$ contains at least one maximal and at least one minimal element.
7. Let $(S, \leq)$ be a finite poset, where $S = \{x_1, \ldots, x_n\}$. Construct the sequence of posets $((S_1, \leq_1), (S_2, \leq_2), \ldots)$ as follows. Let $(S_1, \leq_1) = (S, \leq)$. For $1 \leq i \leq n$, choose $x_{p_i}$ to be the first element of $S_i$ in the sequence $\mathbf{s} = (x_1, \ldots, x_n)$ that is minimal in $(S_i, \leq)$. Define $S_{i+1} = S_i - \{x_{p_i}\}$ and $\leq_{i+1} = \leq_i \cap (S_{i+1} \times S_{i+1})$. Prove that the sequence $(x_{p_1}, \ldots, x_{p_n})$ is a total order on $S$ that extends the partial order $\leq$.
8. Let $S$ be an infinite set and let $(\mathcal{C}, \subseteq)$ be the partially ordered set of its cofinite sets. Prove that for every $U, V \in \mathcal{C}$ both $\sup\{U, V\}$ and $\inf\{U, V\}$ exist.
9. Does the poset of partial functions $(S \rightsquigarrow T, \sqsubseteq)$ introduced in Exercise 2 have a least element?
10. Let $(S, \leq)$ be a poset and let $U$ and $V$ be two subsets of $S$ such that $U \subseteq V$. Prove that if both $\sup U$ and $\sup V$ exist, then $\sup U \leq \sup V$. Prove that if both $\inf U$ and $\inf V$ exist, then $\inf V \leq \inf U$.

11. Prove that the Completeness Axiom of $\mathbb{R}$ implies that for any positive real numbers $x, y$ there exists $n \in \mathbb{N}$ such that $nx > y$ (Archimedes's property of $\mathbb{R}$).
12. Suppose that $S$ and $T$ are subsets of $\mathbb{R}$ that are bounded above. Prove that $S \cup T$ is bounded above and $\sup S \cup T = \max\{\sup S, \sup T\}$.
13. Let $\pi$ and $\sigma$ be two partitions of a finite set $S$. Prove that $|\pi| + |\sigma| \leq |\pi \wedge \sigma| + |\pi \vee \sigma|$.
14. Prove that if $\pi$ is a partition of a set $S$ and $|\pi| = k$, then there are $\binom{k}{2}$ partitions that cover $\pi$.
15. Let $(S, \leq)$ be a poset. Prove that if a chain in $S$ has at most $p$ elements and an antichain has at most $q$ elements, then $|S| \leq pq$.
16. Let $(S, \leq)$ be a poset. Prove that $(S, \leq)$ is a chain if and only if for every subset $T$ of $S$ both $\sup T$ and $\inf T$ exist and $\{\sup T, \inf T\} \subseteq T$.

Let $(S, \leq)$ be a poset. A *realizer* of $(S, \leq)$ is a family of total orders on $S$, $\mathcal{R} = \{\leq_i \mid i \in I\}$ such that

$$\leq = \bigcap \{\leq_i \mid i \in I\}.$$

If $(S, \leq)$ is a finite poset, the *dimension* of $(S, \leq)$ is the smallest size $d$ of a realizer of $(S, \leq)$. The dimension of a finite poset $(S, \leq)$ is denoted by $dim(S, \leq)$.

17. Let $S = \{x_1, \ldots, x_n\}$ be a finite set. Prove that the discrete partial order $\iota_S$ on $S$ has dimension 2.

    **Solution:** Consider the total order $\leq_1 = TO(x_1, \ldots, x_n)$ and its dual $\leq_2 = TO(x_n, \ldots, x_2, x_1)$. Note that $(x, x') \in \leq_1 \cap \leq_2$ if and only if $x = x'$; that is, if and only if $(x, x') \in \iota_S$.
18. Let $(S_n, \leq)$ be the poset whose Hasse diagram is given in Figure 4.10, where $S_n = \{x_1, \ldots, x_n, y_1, \ldots, x_n\}$. This poset was introduced in [43] and is known as the *standard example*. Prove that $dim(S_n, \leq) = n$.

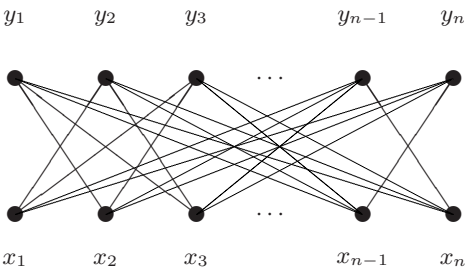



**Fig. 4.10.** The Hasse diagram of the standard example.

19. Consider the poset $(T_{p,m,q}, \leq)$, whose Hasse diagram is given in Figure 4.11. The set $T_{p,m,q}$ consists of three sets of pairwise incomparable elements $\{z_1, \ldots, z_p\}$, $\{u_1, \ldots, u_m\}$, and $\{w_1, \ldots, w_q\}$ such that $z_i < u_j < w_k$ for every $1 \leq i \leq p$, $1 \leq j \leq m$, and $1 \leq k \leq q$. Prove that if at least one of the numbers $p, m, q$ is greater than 1, then

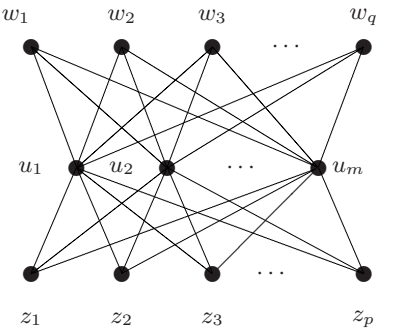



**Fig. 4.11.** Hasse diagram of the poset $T_{m,p,q}$.

$dim(T_{p,m,q}, \leq) = 2$.

20. Prove that the set of partial order relations on a set $S$ is a closure system on the set $S \times S$.
21. Prove that the transitive closure of an acyclic relation is a strict partial order.
22. Let $\mathbf{K}$ be a closure operator on a set $S$. Prove the following statements:
    a) if $U \in \mathcal{C}_{\mathbf{K}}$ and $X \subseteq U \subseteq \mathbf{K}(X)$, then $\mathbf{K}(X) = U$;
    b) $\mathbf{K}(X) \cap \mathbf{K}(Y) \supseteq \mathbf{K}(X \cap Y)$;
    c) $\mathbf{K}(X) \cap \mathbf{K}(Y) \in \mathcal{C}_{\mathbf{K}}$,

    for $X, Y \in \mathcal{P}(S)$.
23. Let $S$ and $T$ be two sets and let $f : S \longrightarrow T$ be a function. Suppose that $\mathbf{K}$ and $\mathbf{L}$ are two closure operators on $S$ and $T$, respectively, such that if $V \in \mathcal{C}_{\mathbf{L}}$, then $f^{-1}(V) \in \mathcal{C}_{\mathbf{K}}$. Prove that, for every $W \in \mathcal{C}_{\mathbf{L}}$ we have $S - f^{-1}(T - W) \in \mathcal{C}_{\mathbf{K}}$.
24. Let $\mathbf{K}$ be a closure operator on a set $S$. For $U \in \mathcal{P}(S)$, define the ***K**-border of the set* $U$ as $\partial_{\mathbf{K}}(U) = \mathbf{K}(U) \cap \mathbf{K}(S - U)$. Let $S$ and $T$ be two sets and let $\mathbf{K}, \mathbf{L}$ be two closure operators on $S$ and $T$, respectively.
    a) Prove that if $f^{-1}(\mathbf{K}(V)) = \mathbf{L}(f^{-1}(V))$ for every $V \in \mathcal{P}(T)$, then $f^{-1}(\partial_{\mathbf{L}}(V)) = \partial_{\mathbf{K}}(f^{-1}(V))$.
    b) Now let $f : S \longrightarrow T$ be a bijection such that both $f^{-1}(\mathbf{K}(V)) = \mathbf{L}(f^{-1}(V))$ for every $V \in \mathcal{P}(T)$ and $f(\mathbf{K}(U)) = \mathbf{L}(f(U))$ for every $U \in \mathcal{P}(S)$. Prove that $\partial_{\mathbf{K}}(f^{-1}(V)) = f^{-1}(\partial_{\mathbf{L}}(V))$ and $\partial_{\mathbf{L}}(f(U)) = f(\partial_{\mathbf{K}}(U))$ for $U \in \mathcal{P}(S)$ and $V \in \mathcal{P}(T)$.

25. Prove that if $\{(S_i, \leq_i) \mid 1 \leq i \leq n\}$ is a family of totally ordered posets, then the lexicographic product $(S_1 \times \cdots \times S_n, \preceq)$ is a total order.
26. Let $(S_1, \leq_1)$ and $(S_2, \leq_2)$ be two posets and let $f : S_1 \longrightarrow S_2$ be a monotonic mapping. Prove that if $S_2$ has a least element 0, then $f^{-1}(0)$ is an order filter of $S_1$, and if $S_2$ has a greatest element 1, then $f^{-1}(1)$ is an order ideal of $S_1$.
27. Let $(S, \leq)$ be a poset. Define the mapping $f_{\leq} : S \longrightarrow \mathcal{P}(S)$ by $f_{\leq}(x) = \{y \in S \mid x < y\}$.
    a) Prove that $f_{\leq}$ is an antimonotonic mapping between the posets $(S, \leq)$ and $(\mathcal{P}(S), \subseteq)$.
    b) If $C$ is a chain in $(S, \leq)$, prove that $f_{\leq}(C)$ is a chain in $(\mathcal{P}(S), \subseteq)$.
    c) Let $(S, \leq)$ and $(S, \leq')$ be two posets defined on the set $S$. Prove that $f_{\leq \cap \leq'}(x) = f_{\leq}(x) \cap f_{\leq'}(x)$ for every $x \in S$.
28. In the proof of Szpilrajn's theorem, we introduced the set of partial orders that extend the partial order "$\leq$". The inclusion between relations defines a partial order on $PORD(S, \leq)$. We saw that the maximal elements of $PORD(S, \leq)$ are total orders on $S$ and that the least element of $PORD(S, \leq)$ is the relation $\leq$ itself.
    Let $(S, \leq)$ be a poset. Prove that there exists a collection of total orders $\{\leq_i \mid i \in I\}$ on $S$ such that $\leq = \bigcap_{i \in I} \leq_i$.

    **Solution:** If $\leq$ is itself a total order, then the desired collection of total orders consists of $\leq$ itself. Suppose therefore that $\leq$ is not total, and let $INC(S, \leq)$ be the set of all pairs of incomparable elements of $(S, \leq)$.

    For each pair $(x, y) \in INC(S, \leq)$, consider the total orders $\leq'_{xy}$ and $\leq''_{xy}$ that extend $\leq$ such that $x \leq'_{xy} y$ and $y \leq'_{xy} x$. Clearly,

$$\leq \subseteq \bigcap \{\leq'_{xy} \cap \leq''_{xy} \mid (x, y) \in INC(S, \leq)\}.$$

    Suppose that $\bigcap \{\leq'_{xy} \cap \leq''_{xy} \mid (x, y) \in INC(S, \leq)\}$ contains a pair of elements $(r, s) INC(S, \leq)$. Then, we have both $r \leq'_{rs} s$ and $r \leq''_{rs} s$. Since $s \leq''_{rs} r$, this would imply $r = s$ by the antisymmetry of $\leq''_{rs}$. This, however, contradicts the incomparability of $(r, s)$ in $(S, \leq)$. Thus, for any pair $(u, v) \in \bigcap \{\leq'_{xy} \cap \leq''_{xy} \mid (x, y) \in INC(S, \leq)\}$, we have $u \leq v$ or $v \leq u$, which shows that

$$\leq = \bigcap \{\leq'_{xy} \cap \leq''_{xy} \mid (x, y) \in INC(S, \leq)\}.$$

29. Let $S$ be a finite set. Prove that the poset $(\mathbf{Seq}(S), \leq_{inf})$, where $\leq_{inf}$ is the partial order introduced in Example 4.6, is locally finite.
30. Let $\zeta : S \times S \longrightarrow \mathbb{R}$ be the Riemann function of a locally finite poset $(S, \leq)$, and let $\zeta^k$ be the product $\zeta * \zeta * \cdots * \zeta$, which contains $k$ *zeta* factors, where $k \in \mathbb{N}$. Prove that:
    a) $\zeta^2(x, y) = |[x, y]|$ if $x \leq y$.
    b) $\zeta^k(x, y)$ gives the number of multichains of length $k$ that can be interpolated between $x$ and $y$.

31. Prove that the Möbius function of the poset $(\mathbb{N}, \delta)$, is given by

$$\mu(n) = \begin{cases} 1 & \text{if } n \text{ is a product of an even number of distinct primes,} \\ -1 & \text{if } n \text{ is a product of an odd number of distinct primes,} \\ 0 & \text{otherwise,} \end{cases}$$

for $n \in \mathbb{N}$.

32. If $\mu$ is the Möbius function of the poset $(\mathbb{N}, \delta)$ prove that

$$\sum\{\mu(m) \mid (m, n) \in \delta\} = \begin{cases} 1 & \text{if } n = 1, \\ 0 & \text{otherwise.} \end{cases}$$

**Solution:** For $n = 1$, the equality is immediate. Suppose that $n > 1$. Only numbers $m$ that are products of distinct prime numbers contribute to the sum $\sum\{\mu(m) \mid (m, n) \in \delta\}$. If $n = p_1^{a_1} \cdots p_r^r$, then this sum equals $\sum_{i=0}^{r} \binom{r}{i}(-1)^i = 0$.

## Bibliographical Comments

There is a vast body of literature dealing with posets and their applications and a substantial number of references that focus on combinatorial study of posets. Among these we mention [135, 136, 128, 138].

# 5

# Lattices and Boolean Algebras

## 5.1 Introduction

Lattices can be defined either as special partially ordered sets or as algebras. In this chapter, we present both definitions and show their equivalence. We study several special classes of lattices: modular and distributive lattices and complete lattices. The last part of the chapter is dedicated to Boolean algebras and Boolean functions and to their applications in data mining.

## 5.2 Lattices as Partially Ordered Sets and Algebras

We begin with a simple algebraic structure.

**Definition 5.1.** *A* semilattice *is a semigroup* $\mathcal{S} = (S, \{*\})$ *such that* $s * s = s$ *and* $s * t = t * s$ *for all* $s, t \in S$.

*In other words,* $\mathcal{S} = (S, \{*\})$ *is a semilattice if "*$*$*" is a commutative and idempotent operation.*

*Example 5.2.* Let $*$ be the binary operation on the set $\mathbb{N}_1$ of positive natural numbers defined by $n * p = \gcd(n, p)$. In Example 2.12, we saw that $*$ is an associative operation. Since $\gcd(n, p) = \gcd(p, n)$ and $\gcd(n, n) = n$ for every $n \in \mathbb{N}$, it follows that $(\mathbb{N}_1, \{*\})$ is indeed a semilattice.

It is easy to see that $(\mathbb{N}_1, \{\mathrm{lcm}\})$ is also a semilattice.

**Theorem 5.3.** *Let* $\mathcal{S} = (S, \{*\})$ *be a semilattice. The relation* $x \leq y$ *defined by* $x = x * y$ *for* $x, y \in S$ *is a partial order on* $S$. *Further,* $\inf\{u, v\}$ *in the partially ordered set* $(S, \leq)$ *exists for all* $u, v \in S$ *and* $u * v = \inf\{u, y\}$.

*Proof.* The idempotency of $*$, $x = x * x$ implies $x \leq x$ for every $x \in S$; that is, the reflexivity of $\leq$.

Suppose that $x \leq y$ and $y \leq x$; that is, $x = x * y$ and $y = y * x$. The commutativity of $*$ implies that $x = y$, so $*$ is antisymmetric.

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_5, © Springer-Verlag London Limited 2008

Now let $x, y, z$ be three elements of $S$ such that $x \leq y$ and $y \leq z$, that is, $x = x * y$ and $y = y * z$. We can write

$$\begin{aligned} x * z &= (x * y) * z \\ &= x * (y * z) \\ &\quad \text{(due to the associativity of } *) \\ &= x * y, \end{aligned}$$

which proves that $x \leq z$. Thus, $\leq$ is transitive, so it is a partial order on $S$.

Let $u$ and $v$ be two arbitrary elements of $S$. Note that $u * v \leq u$ and $v * v \leq u$ because

$$(u * v) * u = u * (u * v) = (u * u) * v = u * v$$

and

$$(u * v) * v = u * (v * v) = u * v.$$

Thus, $u * v$ is a lower bound of the set $\{u, v\}$. Suppose now that $w$ is an arbitrary lower bound of $\{u, v\}$, that is, $w = w * u$ and $w = w * v$. We have $w * (u * v) = (w * u) * v = w * v = w$, which proves that $w \leq u * v$. This allows us to conclude that $u * v$ is indeed the largest lower bound of $\{u, v\}$; that is, $u * v = \inf\{u, v\}$. □

We also need the following converse result.

**Theorem 5.4.** *Let $(S, \leq)$ be a partially ordered set such that $\inf\{u, v\}$ exists for all $u, v \in S$. If $*$ is the operation defined by $u * v = \inf\{u, v\}$ for $u, v \in S$, then $(S, \{*\})$ is a semilattice.*

*Proof.* It is immediate that $*$ is an idempotent and commutative operation. We prove here only its associativity.

Let $t, u, v$ be three elements of $S$ and let $p, q$ be defined by

$$\begin{aligned} p &= \inf\{t, \inf\{u, v\}\}, \\ q &= \inf\{\inf\{t, u\}, v\}. \end{aligned}$$

By the definition of infimum, we have $p \leq t$ and $p \leq \inf\{u, v\}$, so $p \leq u$ and $p \leq v$. Since $p \leq t$ and $p \leq u$, we have $p \leq \inf\{t, u\}$. This inequality together with $p \leq v$ implies $p \leq \inf\{\inf\{t, u\}, v\}$, so $p \leq q$.

By the same definition of infimum, we have $q \leq \inf\{t, u\}$ and $q \geq v$. The first inequality implies $q \leq t$ and $q \leq u$. Thus, $q \leq \inf\{u, v\}$; together with $q \leq t$, these inequalities allow us to write $q \leq \inf\{t, \inf\{u, v\}\} = p$. We conclude that $p = q$, which shows that $*$ is indeed an associative operation. □

The next statement is closely related to the previous theorem.

**Theorem 5.5.** *Let $(S, \leq)$ be a partially ordered set such that $\sup\{u, v\}$ exists for all $u, v \in S$. If $\star$ is the operation defined by $u \star v = \sup\{u, v\}$ for $u, v \in S$, then $(S, \{\star\})$ is a semilattice.*

*Proof.* This statement follows from Theorem 5.4 by duality. □

*Example 5.6.* The partially ordered sets $(P, \leq)$ and $(Q, \leq)$, whose Hasse diagrams are given in Figure 5.1, are semilattices because $\sup\{u, v\}$ exists for any pair of elements in each of these sets. The operation $\star$ is described by the the following table.

| $(P, \star)$ | $a$ | $b$ | $c$ |
|---|---|---|---|
| $a$ | $a$ | $c$ | $c$ |
| $b$ | $c$ | $b$ | $c$ |
| $c$ | $c$ | $c$ | $c$ |

| $(Q, \star)$ | $x$ | $y$ | $z$ |
|---|---|---|---|
| $x$ | $x$ | $y$ | $z$ |
| $y$ | $y$ | $y$ | $z$ |
| $z$ | $z$ | $z$ | $z$ |

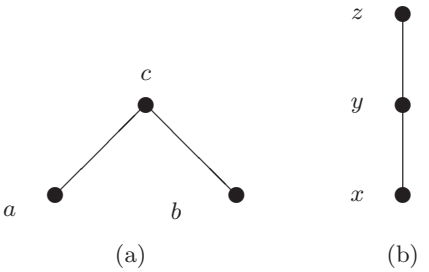



**Fig. 5.1.** Hasse diagrams of the posets $(P, \leq)$ and $(Q, \leq)$.

*Example 5.7.* Let $S$ be a set and let $(\mathbf{Seq}(S), \leq_{pref})$ be the poset introduced in Example 4.6. We will prove that this is a semilattice by verifying that $\inf\{\mathbf{u}, \mathbf{v}\}$ exists for any sequences $\mathbf{u}, \mathbf{v} \in \mathbf{Seq}(S)$.

Note that any two sequences have at least the null sequence $\boldsymbol{\lambda}$ as a common prefix. If $\mathbf{t}$ and $\mathbf{s}$ are common prefixes of $\mathbf{u}$ and $\mathbf{v}$, then either $\mathbf{t}$ is a prefix of $\mathbf{s}$ or vice-versa. Thus, the finite set of common prefixes of $\mathbf{u}$ and $\mathbf{v}$ is totally ordered by "$\leq_{pref}$" and therefore it has a largest element $\mathbf{z}$. The sequence $\mathbf{z}$ is the longest common prefix of the sequences $\mathbf{u}$ and $\mathbf{v}$. It is clear that $\mathbf{z} = \inf\{\mathbf{u}, \mathbf{v}\}$ in the poset $(\mathbf{Seq}(S), \leq_{pref})$.

We will denote the result of the semilattice operation introduced here, which associates with $\mathbf{u}$ and $\mathbf{v}$ their longest common prefix, by $lcp(\mathbf{u}, \mathbf{v})$.

The associativity of this operation can be written as

$$lcp(\mathbf{u}, lcp(\mathbf{v}, \mathbf{w})) = lcp(lcp(\mathbf{u}, \mathbf{v}), \mathbf{w}) \tag{5.1}$$

for all sequences $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{Seq}(S)$.

A useful property of the semilattice $(\mathbf{Seq}(S), lcp)$ introduced in Example 5.7 is given next.

**Theorem 5.8.** *Let $S$ be a set and let $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ be three sequences in* $\mathbf{Seq}(S)$. *Then, at most two of the sequences $lcp(\boldsymbol{u}, \boldsymbol{v})$, $lcp(\boldsymbol{v}, \boldsymbol{w})$, $lcp(\boldsymbol{w}, \boldsymbol{u})$ are distinct. The common value of two of these sequences is a prefix of the third sequence.*

*Proof.* Let $\mathbf{t} = lcp(\mathbf{u}, \mathbf{v})$, $\mathbf{r} = lcp(\mathbf{v}, \mathbf{w})$, and $\mathbf{s} = lcp(\mathbf{w}, \mathbf{u})$. Note that any two of the sequences $\mathbf{t}, \mathbf{r}, \mathbf{s}$ are prefixes of the same sequence. Therefore, they form a chain in the poset $(\mathbf{Seq}(S), \leq_{pref})$.

Suppose, for example, that $\mathbf{t} \leq_{pref} \mathbf{r} \leq_{pref} \mathbf{s}$. Observe that $\mathbf{r}$ is a prefix of $\mathbf{v}$ because it is a prefix of $\mathbf{s}$. Thus, $\mathbf{r}$ is a prefix of both $\mathbf{u}$ and $\mathbf{v}$. Since $\mathbf{t}$ is the longest common prefix of $\mathbf{u}$ and $\mathbf{v}$, it follows that $\mathbf{r}$ is a prefix of $\mathbf{t}$, so $\mathbf{r} = \mathbf{t}$.

The remaining five cases that correspond to the remaining permutation of the sequences $\mathbf{t}, \mathbf{r}$, and $\mathbf{s}$ can be treated in a similar manner. □

Theorems 5.4 and 5.5 show that, in principle, a partial order relation on a set $S$ may induce two semilattice structures on $S$. Traditionally, the semilattice $(S, *)$ has been referred to as the *meet semilattice*, while $(S, \star)$ is called the *join semilattice*, and the operations "$*$" and "$\star$" are denoted by "$\wedge$" and "$\vee$", respectively. This is a notation that we will use from now on.

**Definition 5.9.** *Let $\mathcal{S}_1 = (S_1, \{\wedge\})$ and $\mathcal{S}_2 = (S_2, \{\wedge\})$ be two semilattices. A* morphism *$h$ from $\mathcal{S}_1$ to $\mathcal{S}_2$ is a function $h : S_1 \longrightarrow S_2$ such that $h(x \wedge y) = h(x) \wedge h(y)$ for $x, y \in S_1$.*

*The semilattices $\mathcal{S}_1$ and $\mathcal{S}_2$ are* isomorphic *if there exist two bijective morphisms $h : S_1 \longrightarrow S_2$ and $h' : S_2 \longrightarrow S_1$ that are inverse to each other.*

A semilattice morphism is a monotonic function between the partially ordered sets $(S_1, \leq)$ and $(S_2, \leq)$. Indeed, suppose that $x, y \in S_1$ such that $x \leq y$, which is equivalent to $x = x \wedge y$. Since $h$ is a morphism, we have $h(x) = h(x) \wedge h(y)$, so $h(x) \leq h(y)$. The converse is not true; a monotonic function between the posets $(S_1, \leq)$ and $(S_2, \leq)$ is not necessarily a semilattice morphism, as the next example shows.

*Example 5.10.* Let $(P, \{\star\})$ and $(Q, \{\star\})$ be the semilattices defined in Example 5.6. The function $f : P \longrightarrow Q$ given by $f(a) = x$, $f(b) = y$, and $f(c) = z$ is clearly monotonic, and it is even a bijection. However, it fails to be a semilattice morphism because $f(a \star b) = f(c) = z$, while $f(a) \star f(b) = x \star y = y \neq z$.

However, we have the following theorem.

**Theorem 5.11.** *Let $\mathcal{S}_1 = (S_1, \{\wedge\})$ and $\mathcal{S}_2 = (S_2, \{\wedge\})$ be two semilattices. $\mathcal{S}_1$ and $\mathcal{S}_2$ are isomorphic if and only if there exists a bijection $h : S_1 \longrightarrow S_2$ such that both $h$ and $h^{-1}$ are monotonic.*

*Proof.* Suppose that $h : S_1 \longrightarrow S_2$ is a bijection such that both $h$ and $h^{-1}$ are monotonic functions, and let $x$ and $y$ be two elements of $S_1$. Since $x \wedge y \leq x$ and $x \wedge y \leq y$, we have $h(x \wedge y) \leq h(x)$ and $h(x \wedge y) \leq h(y)$, so $h(x \wedge y) \leq h(x) \wedge h(y)$. We further prove that $h(x \wedge y)$ is the infimum of $h(x)$ and $h(y)$.

Let $u \in S_2$ such that $u \leq h(x) \wedge h(y)$, so $u \leq h(x)$ and $u \leq h(y)$. Equivalently, we have $h^{-1}(u) \leq x$ and $h^{-1}(u) \leq y$, which implies $h^{-1}(u) \leq x \wedge y$. Therefore, $u \leq h(x \wedge y)$, which allows us to conclude that $h(x) \wedge h(y) = \inf\{h(x), h(y)\} = h(x \wedge y)$, so $h$ is indeed a morphism. Similarly, one can prove that $h^{-1}$ is also a morphism, so $\mathcal{S}_1$ and $\mathcal{S}_2$ are isomorphic.

Conversely, if $\mathcal{S}_1$ and $\mathcal{S}_2$ are isomorphic and $h : S_1 \longrightarrow S_2$ and $h' : S_2 \longrightarrow S_1$ are morphisms that are inverse to each other, then they are clearly inverse monotonic mapping. □

A structure that combines the properties of join and meet semilattices is introduced next.

**Definition 5.12.** *A* lattice *is an algebra of type* $(2, 2)$*, that is, an algebra* $\mathcal{L} = (L, \{\wedge, \vee\})$ *such that* $\wedge$ *and* $\vee$ *are both idempotent, commutative, and associative operations and the equalities (known as* absorption laws*)*

$$x \vee (x \wedge y) = x, x \wedge (x \vee y) = x$$

*are satisfied for every* $x, y \in L$*.*

Observe that if $(L, \{\wedge, \vee\})$ is a lattice, then both $(L, \{\wedge\})$ and $(L, \{\vee\})$ are semilattices. Thus, by Theorem 5.3, both operations induce partial order relations on $L$. Let us denote these operations temporarily by "$\leq$" and "$\leq'$", respectively. In other words, we have $x \leq y$ if $x = x \wedge y$ and $u \leq' v$ if $u = u \vee v$.

The absorption laws that link together the operations $\wedge$ and $\vee$ imply that the two partial orders are dual to each other. Indeed, suppose that $x \leq y$, that is, $x = x \wedge y$. Then, since $y \vee x = y \vee (y \wedge x) = y$, we have $y \leq' x$. We usually use the partial order $\leq$ on the lattice $(L, \{\wedge, \vee\})$.

If $(L, \{\wedge, \vee\})$ is a lattice, then for every finite, nonempty subset $K$ of $L$, $\inf K$ and $\sup K$ exist, as it can be shown by induction on $n = |K|$, where $n \geq 1$ (see Exercise 2). Moreover, if $K = \{x_1, \ldots, x_n\}$, then

$$\inf K = x_1 \wedge x_2 \wedge \cdots \wedge x_n,$$
$$\sup K = x_1 \vee x_2 \vee \cdots \vee x_n.$$

*Example 5.13.* Let $S$ be a set. The algebra $(\mathcal{P}(S), \{\cap, \cup\})$ is a lattice. Also, if $\mathcal{P}_f(S)$ is the set of all finite subsets of $S$, then $(\mathcal{P}_f(S), \{\cap, \cup\})$ is also a lattice.

*Example 5.14.* The posets $M_5$ and $N_5$ from Example 4.78 are both lattices. Indeed, the operations $\wedge$ and $\vee$ for the first poset is given by the following table.

| $(M_5, \wedge)$ | 0 | $a$ | $b$ | $c$ | 1 |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| $a$ | 0 | $a$ | 0 | 0 | $a$ |
| $b$ | 0 | 0 | $b$ | 0 | $b$ |
| $c$ | 0 | 0 | 0 | $c$ | $c$ |
| 1 | 0 | $a$ | $b$ | $c$ | 1 |

| $(M_5, \vee)$ | 0 | $a$ | $b$ | $c$ | 1 |
|---|---|---|---|---|---|
| 0 | 0 | $a$ | $b$ | $c$ | 1 |
| $a$ | $a$ | $a$ | 1 | 1 | 1 |
| $b$ | $b$ | 1 | $b$ | 1 | 1 |
| $c$ | $c$ | 1 | 1 | $c$ | 1 |
| 1 | 1 | 1 | $b$ | 1 | 1 |

The similar operations for $N_5$ are given next.

| $(N_5, \wedge)$ | $0$ | $x$ | $y$ | $z$ | $1$ |
|---|---|---|---|---|---|
| $0$ | $0$ | $0$ | $0$ | $0$ | $0$ |
| $x$ | $0$ | $x$ | $x$ | $0$ | $x$ |
| $y$ | $0$ | $x$ | $y$ | $0$ | $y$ |
| $z$ | $0$ | $0$ | $0$ | $z$ | $z$ |
| $1$ | $0$ | $x$ | $y$ | $1$ | $1$ |

| $(N_5, \vee)$ | $0$ | $x$ | $y$ | $z$ | $1$ |
|---|---|---|---|---|---|
| $0$ | $0$ | $x$ | $y$ | $z$ | $1$ |
| $x$ | $x$ | $x$ | $y$ | $1$ | $1$ |
| $y$ | $y$ | $y$ | $y$ | $1$ | $1$ |
| $z$ | $z$ | $1$ | $1$ | $z$ | $1$ |
| $1$ | $1$ | $1$ | $1$ | $1$ | $1$ |

*Example 5.15.* The poset of partitions of a finite set $(PART(S), \leq)$ introduced in Example 4.3 is a lattice. Indeed, we saw in Section 4.8 that for every two partitions $\pi, \sigma$, both $\inf\{\pi, \sigma\}$ and $\sup\{\pi, \sigma\}$ exist.

*Example 5.16.* Consider the set $\mathbb{N} \times \mathbb{N}$ and the partial order $\preceq$ on this set defined by $(p, q) \preceq (m, n)$ if $p \leq m$ and $q \leq n$. Then $\inf\{(u, v), (x, y)\} = (\min\{u, x\}, \min\{v, y\})$ and $\sup\{(u, v), (x, y)\} = (\max\{u, x\}, \max\{v, y\})$.

**Theorem 5.17.** *Let $(L, \{\wedge, \vee\})$ be a lattice. If $x \leq y$ and $u \leq v$, then $x \wedge u \leq y \wedge v$ and $x \vee u \leq y \vee v$ (compatibility of the lattice operations with the partial order).*

*Proof.* Note that $x \leq y$ is equivalent to $x = x \wedge y$ and to $y = x \vee y$. Similarly, $u \leq v$ is equivalent to $u = u \wedge v$ and to $v = u \vee v$. Therefore, we can write

$$(x \wedge u) \wedge (y \wedge v) = (x \wedge y) \wedge (u \wedge v) = x \wedge u,$$

so $x \wedge u \leq y \wedge v$. The proof of the second inequality is similar. □

Let $(L, \{\wedge, \vee\})$ be a lattice. If the poset $(L, \leq)$ has the largest element 1, then we have $1 \wedge x = x \wedge 1 = x$ and $1 \vee x = x \vee 1 = x$. If the poset has the least element 0, then $0 \wedge x = x \wedge 0 = 0$ and $0 \vee x = x \vee 0 = x$. In other words, if a lattice has a largest element 1, then 1 is a unit with respect to the $\wedge$ operation; similarly, if the least element exists, then it plays the role of a unit with respect to $\vee$.

Let $K$ and $H$ be two finite subsets of $L$, where $(L, \{\wedge, \vee\})$ is a lattice. If $K \subseteq H$, then it is easy to see that $\sup K \leq \sup H$ and that $\inf K \geq \inf H$. Since $\emptyset \subseteq H$ for every set $H$, by choosing $H = \{x\}$ for some $x \in L$, it is clear that if a lattice has the least element 0 and the greatest element 1, then we can define $\sup \emptyset = 0$ and $\inf \emptyset = 1$.

**Definition 5.18.** *A lattice $(L, \{\wedge, \vee\})$ is* bounded *if the poset $(L, \leq)$ has the least element and the greatest element* 1*.*

If a lattice $(L, \{\wedge, \vee\})$ is bounded, then every finite subset of $L$ (including the empty set) is bounded.

If $\mathfrak{L} = (L, \{\wedge, \vee\})$ is a finite lattice, then $\mathfrak{L}$ is bounded. Indeed, since both $\sup L$ and $\inf L$ exist, it follows that $\sup L$ is the greatest element and $\inf L$ is the least element of $\mathfrak{L}$, respectively.

**Definition 5.19.** *Let $\mathfrak{L}_1 = (L_1, \wedge, \vee)$ and $\mathfrak{L}_2 = (L_2, \wedge, \vee)$ be two lattices. A* morphism *$h$ from $\mathfrak{L}_1$ to $\mathfrak{L}_2$ is a function $h : L_1 \longrightarrow L_2$ such that $h(x \wedge y) = h(x) \wedge h(y)$ and $h(x \vee y) = h(x) \vee h(y)$ for every $x, y \in L_1$.*

*A* lattice isomorphism *is a bijective lattice morphism.*

A counterpart of Theorem 5.11 characterizes isomorphic lattices.

**Theorem 5.20.** *Let $\mathfrak{L}_1 = (L_1, \{\wedge, \vee\})$ and $\mathfrak{L}_2 = (L_2, \{\wedge, \vee\})$ be two lattices. $\mathfrak{L}_1$ and $\mathfrak{L}_2$ are isomorphic if and only if there exists a bijection $h : L_1 \longrightarrow L_2$ such that both $h$ and $h^{-1}$ are monotonic.*

*Proof.* The proof is similar to the proof of Theorem 5.11; we leave the argument to the reader as an exercise. □

**Definition 5.21.** *Let $\mathfrak{L} = (L, \{\wedge, \vee\})$ be a lattice. A* sublattice *of $\mathfrak{L}$ is a subset $K$ of $L$ that is closed with respect to the lattice operations. In other words, for every $x, y \in K$, we have both $x \wedge y \in K$ and $x \vee y \in K$.*

Note that if $K$ is a sublattice of $\mathfrak{L}$, then the pair $\mathcal{K} = (K, \{\wedge, \vee\})$ is itself a lattice. We use the term "sublattice" to designate both the set $K$ and the lattice $\mathcal{K}$ when there is no risk of confusion. For example, if $K$ and $K'$ are two sublattices of a lattice $\mathfrak{L}$ and $f : K \longrightarrow K'$ is a morphism between the lattices $\mathcal{K} = (K, \{\wedge, \vee\})$ and $\mathcal{K}' = (K', \{\wedge, \vee\})$, we designate $f$ as a morphism between $K$ and $K'$.

*Example 5.22.* Let $\mathfrak{L} = (L, \{\wedge, \vee\})$ be a lattice and let $a$ and $b$ be a pair of elements of $L$. The *interval* $[a, b]$ is the set

$$\{x \in L \mid a \leq x \leq b\}.$$

Clearly, an interval $[a, b]$ is nonempty if and only if $a \leq b$. Each such set is a sublattice. Indeed, if $[a, b] = \emptyset$, then $\emptyset$ is clearly a sublattice.

Suppose that $x, y \in [a, b]$; that is, $a \leq x \leq b$ and $a \leq y \leq b$. Due to the compatibility of the lattice operations with the partial order, we obtain immediately $a \leq x \wedge y \leq b$ and $a \leq x \vee y \leq b$, so $[a, b]$ is a sublattice in all cases.

*Example 5.23.* Let $[a, b]$ be a nonempty interval of a lattice $\mathfrak{L} = (L, \{\wedge, \vee\})$. The function $h : L \longrightarrow [a, b]$ defined by $h(x) = (x \vee a) \wedge b$ is a surjective morphism between $\mathfrak{L}$ and the lattice $([a, b], \{\wedge, \vee\})$ because

$$\begin{aligned} h(x \wedge y) &= ((x \wedge y) \vee a) \wedge b \\ &= ((x \vee a) \wedge (y \vee a)) \wedge b \\ &= ((x \vee a) \wedge b) \wedge ((y \vee a)) \wedge b)) \\ &= h(x) \wedge h(y) \end{aligned}$$

and

$$\begin{aligned} h(x \vee y) &= ((x \vee y) \vee a) \wedge b \\ &= ((x \vee a) \vee (y \vee a)) \wedge b \\ &= ((x \vee a) \wedge b) \vee ((y \vee a)) \wedge b)) \\ &= h(x) \vee h(y) \end{aligned}$$

for $x, y \in B$.

The elements $a$ and $b$ are invariant under $h$. Indeed, we have $h(a) = a$ because $h(a) = (a \vee a) \wedge b = a \wedge b = a$ and $h(b) = (b \vee a) \wedge b = b$ by absorption. Moreover, this property is shared by every member of the interval $[a, b]$ because we can write

$$\begin{aligned} h(h(x)) &= h((x \vee a) \wedge b) = h(x \vee a) \wedge h(b) \\ &= (h(x) \vee h(a)) \wedge h(b) = (h(x) \vee a) \wedge b \\ &= (((x \vee a) \wedge b) \vee a) \wedge b \\ &= (x \vee a) \wedge (b \vee a) \wedge b \\ &= (x \vee a) \wedge b = h(x) \end{aligned}$$

for $x \in B$. We refer to $h$ as the *projection* of $L$ on the interval $[a, b]$.

## 5.3 Special Classes of Lattices

Let $(L, \{\wedge, \vee\})$ be a lattice and let $u, v, w$ be three members of $L$ such that $u \leq w$. Since $u \leq u \vee v$ and $u \leq w$, it follows that

$$u \leq (u \vee v) \wedge w. \tag{5.2}$$

Starting from the inequalities $v \wedge w \leq v \leq u \vee v$ and $v \wedge w \leq w$, we have also

$$v \wedge w \leq (u \vee v) \wedge w. \tag{5.3}$$

Combining Inequalities (5.2) and (5.3) yields the inequality

$$u \vee (v \wedge w) \leq (u \vee v) \wedge w, \tag{5.4}$$

which is satisfied whenever $u \leq w$. This inequality is known as the *submodular inequality*.

An important class of lattices is obtained when we replace the submodular inequality (satisfied by every lattice) with an equality, as follows.

**Definition 5.24.** *A lattice* $(L, \{\wedge, \vee\})$ *is* modular *if, for every* $u, v, w \in L$, $u \leq w$ *implies*

$$u \vee (v \wedge w) = (u \vee v) \wedge w. \tag{5.5}$$

Observe that if $u = w$, Equality (5.5) holds in every lattice. Therefore, it is sufficient to require that $u < w$ implies $u \vee (v \wedge w) = (u \vee v) \wedge w$ for all $u, v, w \in L$ to ensure modularity.

*Example 5.25.* The lattice $M_5$ introduced in Example 5.14 is modular. Indeed, suppose that $x < z$. If $u = 0$ and $w = 1$, it is easy to see that Equality (5.5) is verified. Suppose, for example, that $u = a$ and $w = 1$. Then, $u \vee (v \wedge w) = a \vee (v \wedge 1) = a \vee v$ and $(u \vee v) \wedge w = (a \vee v) \wedge 1 = a \vee v$ for every $v \in \{0, 1, a, b, c\}$. The remaining cases can be analyzed similarly.

On the other hand, the lattice $N_5$ introduced in the same example is not modular because we have $x < y$, $x \vee (z \wedge y) = x \vee 0 = x$, and $(x \vee z) \wedge y = 1 \wedge y = y \neq x$.

The special role played by $N_5$ is described next.

**Theorem 5.26.** *A lattice $\mathfrak{L} = (L, \{\wedge, \vee\})$ is modular if and only if it does not contain a sublattice isomorphic to $N_5$.*

*Proof.* Suppose $\mathfrak{L}$ contains a sublattice $K = \{t_0, t_1, t_2, t_3, t_4\}$ isomorphic to $N_5$, and let $f : K \longrightarrow N_5$ be an isomorphism. Suppose that $f(t_0) = 0$, $f(t_1) = x$, $f(t_2) = y$, $f(t_3) = z$, and $f(t_4) = 1$. Also, let $g : N_5 \longrightarrow K$ be the inverse isomorphism.

Since $x < y$, $g(x) = t_1$, and $g(y) = t_2$, we have $t_1 < t_2$. On the other hand, $t_1 \vee (t_3 \wedge t_2) = g(x) \vee (g(z) \wedge g(y)) = g(x \vee (z \wedge y)) = g(x) = t_1$ and $(t_1 \vee t_3) \wedge t_2 = (g(x) \vee g(z)) \wedge g(y) = g((x \vee z) \wedge y) = g(y) = t_2 \neq t_1$, which shows that $\mathfrak{L}$ is not modular.

Conversely, suppose that $\mathfrak{L} = (L, \{\wedge, \vee\})$ is not modular. Then, there exist three members of $L$ – $u, v, w$ – such that $u < w$ and $u \vee (v \wedge w) < (u \vee v) \wedge w$ because $\mathfrak{L}$ still satisfies the submodular inequality. Observe that the elements $t_0, \ldots, t_4$ given by:

$$\begin{aligned} t_0 &= v \wedge w, \\ t_1 &= u \vee (v \wedge w), \\ t_2 &= (u \vee v) \wedge w, \\ t_3 &= v, \\ t_4 &= (u \vee v) \wedge w \end{aligned}$$

form a sublattice isomorphic to $N_5$. □

An important property of modular lattices relates intervals of the form $[a \wedge b, a]$ and $[b, a \vee b]$ for any $a, b \in L$.

**Theorem 5.27.** *Let $\mathfrak{L} = (L, \{\wedge, \vee\})$ be a modular lattice and let $a$ and $b$ be two elements. The mappings $\phi : [a \wedge b, a] \longrightarrow [b, a \vee b]$ and $\psi : [b, a \vee b] \longrightarrow [a \wedge b, a]$ defined by $\phi(x) = x \vee b$ and $\psi(y) = y \wedge a$ for $x \in [a \wedge b, b]$ and $y \in [a, a \vee b]$ are inverse monotonic mappings between the sublattices $[a \wedge b, b]$ and $[a, a \vee b]$.*

*Proof.* Note that, for $a \wedge b \leq x \leq a$, we have

$$\begin{aligned}(x \vee b) \wedge a &= x \vee (b \wedge a) \\ &\quad \text{(since } \mathfrak{L} \text{ is modular)} \\ &= x \text{(because } a \wedge b \leq x\text{)}.\end{aligned}$$

Thus, $\psi(\phi(x)) = x$ for every $x \in [a \wedge b, a]$. Similarly, one can prove that $\phi(\psi(y)) = y$ for every $y \in [b, a \vee b]$, which shows that $\phi$ and $\psi$ are inverse to each other. The monotonicity is immediate. □

**Corollary 5.28.** *Let $\mathfrak{L} = (L, \{\wedge, \vee\})$ be a modular lattice and let $a$ and $b$ be two elements such that $a$ and $b$ cover $a \wedge b$. Then $a \vee b$ covers both $a$ and $b$.*

*Proof.* Since $a$ covers $a \wedge b$, the interval $[a \wedge b, a]$ consists of two elements. Therefore, by Theorem 5.27, the interval $[b, a \vee b]$ also consists of two elements, so $a \vee b$ covers $b$. A similar argument shows that $a \vee b$ covers $a$ (starting from the fact that $b$ covers $a \wedge b$). □

The property of modular lattices described in Corollary 5.28 allows us to introduce a generalization of the class of modular lattices.

**Definition 5.29.** *A lattice $\mathfrak{L} = (L, \{\wedge, \vee\})$ is* semimodular*, if for every $a, b \in L$ such that both cover $a \wedge b$, the same elements are covered by $a \vee b$.*

Clearly, every modular lattice is semimodular. The converse is not true, as the next example shows.

*Example 5.30.* Let $(PART(S), \{\wedge, \vee\})$ the lattice of partitions of the set $S = \{1, 2, 3, 4\}$, whose Hasse diagram is shown in Figure 4.3. By Theorem 4.53, a partition $\sigma$ covers the partition $\pi$ if and only if there exists a block $C$ of $\sigma$ that is the union of two blocks $B$ and $B'$ of $\pi$, and every block of $\sigma$ that is distinct of $C$ is a block of $\pi$. Thus, it is easy to verify that this lattice is indeed semimodular.

To show that $(PART(\{1, 2, 3, 4\}), \{\wedge, \vee\})$ is not modular, consider the partitions

$$\begin{aligned}\pi_1 &= \{12, 3, 4\}, \\ \pi_2 &= \{123, 4\}, \\ \pi_3 &= \{14, 2, 3\}.\end{aligned}$$

It is easy to see that the sublattice $\alpha_S, \pi_1, \pi_2, \pi_3, \omega_S$ is isomorphic to $N_5$ and therefore the lattice is not modular.

A more general statement follows.

**Theorem 5.31.** *The partition lattice $(PART(S), \{\wedge, \vee\})$ of a nonempty set is semimodular.*

*Proof.* Let $\pi, \sigma \in PART(S)$ be two partitions such that both cover $\pi \wedge \sigma$. By Theorem 4.53, both $\pi$ and $\sigma$ are obtained from $\pi \wedge \sigma$ by fusing two blocks of this partition. If $\pi \wedge \sigma = \{B_1, \ldots, B_n\}$, then there exist three blocks of $\pi \wedge \sigma$, $B_p, B_q, B_r$, such that $\pi$ is obtained by fusing $B_p$ and $B_q$, and $\sigma$ is obtained by fusing $B_q$ and $B_r$. To simplify the argument we can assume without loss of generality that $p = 1$, $q = 2$, and $r = 3$.

The graph $G_{\pi,\sigma}$ of the partitions $\pi$ and $\sigma$ is given in Figure 5.2. The blocks of the partition $\pi \vee \sigma$ correspond to the connected components of the graph $G_{\pi,\sigma}$, so $\pi \vee \sigma = \{B_1 \cup B_2 \cup B_3, \ldots, B_n\}$, which covers both $\pi$ and $\sigma$. Thus, $(PART(S), \{\wedge, \vee\})$ is semimodular.

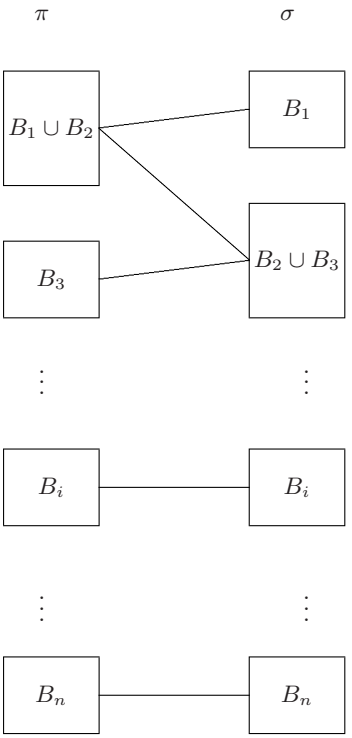



**Fig. 5.2.** The graph $G_{\pi,\sigma}$ of $\pi$ and $\sigma$.

□

*Example 5.32.* Let $\mathfrak{L} = (\{0, a, b, c, d, e, 1\}, \{\wedge, \vee\})$ be the lattice whose Hasse diagram is shown in Figure 5.3. This is a semimodular lattice that is not modular. Indeed, we have $a \leq c$ but $(a \vee e) \wedge c = c$, while $a \vee (e \wedge c) = a \neq c$.

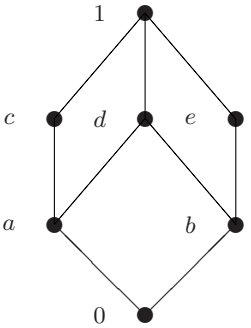



**Fig. 5.3.** Hasse diagram of lattice $\mathcal{L} = (\{0, a, b, c, d, e, 1\}, \{\wedge, \vee\})$.

Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a lattice and let $x, y, z$ be three elements of $L$. We have the inequalities

$$x \wedge (y \vee z) \geq (x \wedge y) \vee (x \wedge z), \tag{5.6}$$
$$x \vee (y \wedge z) \leq (x \vee y) \wedge (x \vee z). \tag{5.7}$$

Indeed, note that $x \geq x \wedge y$ and $x \geq x \wedge z$, so $x \geq (x \wedge y) \vee (x \wedge z)$. Also, $(y \vee z) \geq (x \wedge y)$ and $(y \vee z) \geq (x \wedge z)$, which implies $(y \vee z) \geq (x \wedge y) \vee (x \wedge z)$. Therefore, we conclude that $x \wedge (y \vee z) \geq (x \wedge y) \vee (x \wedge z)$. The argument for the second inequality is similar. We refer to Inequalities (5.6) and (5.7) as the *subdistributive inequalities*.

The existence of subdistributive inequalities satisfied by every lattice serves as an introduction to a very important class of lattices, which we define next.

**Definition 5.33.** *A lattice* $(L, \{\wedge, \vee\})$ *is* distributive *if*

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z),$$
$$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z),$$

*for* $x, y, z \in L$.

In fact, it is sufficient that only one of the equalities of Theorem 5.33 be satisfied to ensure distributivity. Suppose, for example, that $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ for $x, y, z \in L$. We have

$$\begin{aligned}
x \vee (y \wedge z) &= (x \vee (x \wedge z)) \vee (y \wedge z) \\
&\quad \text{(by absorption)} \\
&= x \vee ((x \wedge z) \vee (y \wedge z)) \\
&\quad \text{(by the associativity of } \vee) \\
&= x \vee ((z \wedge x) \vee (z \wedge y)) \\
&\quad \text{(by the commutativity of } \wedge) \\
&= x \vee (z \wedge (x \vee y)) \\
&\quad \text{(by the first distributivity equality)} \\
&= x \vee ((x \vee y) \wedge z) \\
&\quad \text{(by the commutativity of } \wedge) \\
&= (x \wedge (x \vee y)) \vee ((x \vee y) \wedge z) \\
&\quad \text{(by absorption)} \\
&= ((x \vee y) \wedge x) \vee ((x \vee y) \wedge z) \\
&\quad \text{(by the commutativity of } \wedge) \\
&= (x \vee y) \wedge (x \vee z), \\
&\quad \text{(by the first distributivity equality),}
\end{aligned}$$

which is the second distributivity law. In a similar manner, one could show that the second distributivity law implies the first law.

**Theorem 5.34.** *Every distributive lattice is modular.*

*Proof.* Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a distributive lattice. Suppose that $u \leq w$. Applying the distributivity, we can write

$$\begin{aligned}
u \vee (v \wedge w) &= (u \vee v) \wedge (u \vee w) \\
&= (u \vee v) \wedge w, \\
&\quad \text{(because } u \leq w),
\end{aligned}$$

which shows that $\mathcal{L}$ is modular. □

We saw that the lattice $N_5$ is not modular and therefore is not distributive. The lattice $M_5$ is modular (as we have shown in Example 5.25) but not distributive. Indeed, note that

$$a \vee (b \wedge c) = a \vee 0 = a$$

and

$$(a \vee b) \wedge (a \vee c) = 1 \wedge 1 \neq 0.$$

It is easy to see that every sublattice of a distributive lattice is also distributive. Thus, a distributive lattice may not contain sublattices isomorphic to $M_5$ or $N_5$. This allows the formulation of a statement for distributive lattices that is similar to Theorem 5.26.

**Theorem 5.35.** *A lattice $\mathcal{L} = (L, \{\wedge, \vee\})$ is distributive if and only if it does not contain a sublattice isomorphic to $M_5$ or $N_5$.*

*Proof.* The necessity of this condition is clear, so we need to prove only that it is sufficient.

Let $\mathcal{L}$ be a lattice that is not distributive. Then, $\mathcal{L}$ may or may not be modular. If $\mathcal{L}$ is not modular, then by Theorem 5.26 it contains a sublattice isomorphic to $N_5$. Therefore, we need to consider only the case where $\mathcal{L}$ is modular but not distributive. We show in this case that $\mathcal{L}$ contains a sublattice that is isomorphic to $M_5$.

The nondistributivity of $\mathcal{L}$ implies the existence of $x, y, z \in L$ such that

$$x \wedge (y \vee z) > (x \wedge y) \vee (x \wedge z), \tag{5.8}$$
$$x \vee (y \wedge z) < (x \vee y) \wedge (x \vee z). \tag{5.9}$$

Let $u$ and $v$ be defined by

$$u = (x \vee y) \wedge (y \vee z) \wedge (z \vee x)$$
$$v = (x \wedge y) \vee (y \wedge z) \vee (z \wedge x).$$

We first prove that $v < u$.

Note that

$$\begin{aligned}
&x \wedge u \\
&= x \wedge ((x \vee y) \wedge (y \vee z) \wedge (z \vee x)) \text{ (by the definition of } u) \\
&= (x \wedge (x \vee y)) \wedge (y \vee z) \wedge (z \vee x) \text{ (by the associativity of } \wedge) \\
&= x \wedge (y \vee z) \wedge (z \vee x) \text{ (by absorption)} \\
&= x \wedge (z \vee x) \wedge (y \vee z) \text{ (by associativity and commutativity of } \wedge) \\
&= x \wedge (x \vee z) \wedge (y \vee z) \text{ (by commutativity of } \vee) \\
&= x \wedge (y \vee z) \text{ (by absorption).}
\end{aligned}$$

Also,

$$\begin{aligned}
&x \vee v \\
&= x \vee ((x \wedge y) \vee (y \wedge z) \vee (z \wedge x)) \text{ (by the definition of } v) \\
&= x \vee ((x \wedge y) \vee (x \wedge z) \vee (y \wedge z)) \text{ (by associativity and commutativity)} \\
&= ((x \wedge y) \vee (x \wedge z)) \vee (x \wedge (y \wedge z) \\
&\quad \text{(by modularity since } (x \wedge y) \vee (x \wedge z) \leq x) \\
&= (x \wedge y) \vee (x \wedge z) \text{(because } x \wedge y \wedge z \leq (x \wedge y) \vee (x \wedge z)).
\end{aligned}$$

Thus, by Inequality (5.8), we have $x \vee v < x \vee u$, which clearly implies $v < u$.

Consider now the projections $x_1, y_1, z_1$ of $x, y, z$ on the interval $[v, u]$ given by

$$x_1 = (x \wedge u) \vee v, y_1 = (y \wedge u) \vee v, z_1 = (z \wedge u) \vee v.$$

It is clear that $v \leq x_1, y_1, z_1 \leq u$. We shall prove that $\{u, x_1, y_1, z_1, v\}$ is a sublattice isomorphic to $M_5$ by showing that

$$x_1 \wedge y_1 = y_1 \wedge z_1 = z_1 \wedge x_1 = v \text{ and } x_1 \vee y_1 = y_1 \vee z_1 = z_1 \vee x_1 = u.$$

We have

$$\begin{aligned}
&x_1 \wedge y_1 \\
&= ((x \wedge u) \vee v) \wedge ((y \wedge u) \vee v) \text{ (by the definition of } x_1 \text{ and } y_1) \\
&= ((x \wedge u) \wedge ((y \wedge u) \vee v)) \vee v \text{ (by modularity since } v \leq (y \wedge u) \vee v) \\
&= ((x \wedge u) \wedge ((y \vee v) \wedge u)) \vee v \text{ (by modularity since } v \leq u) \\
&= ((x \wedge u) \wedge u \wedge (y \vee v)) \vee v \text{ (by associativity and commutativity of } \wedge) \\
&= ((x \wedge u) \wedge (y \vee v)) \vee v \text{ (by absorption)} \\
&= (x \wedge (y \vee z) \wedge (y \vee (x \wedge z))) \vee v \text{ (because } x \wedge u = x \wedge (y \vee z) \text{ and} \\
&\quad y \vee v = y \vee (x \wedge z)) \\
&= (x \wedge (y \vee ((y \vee z) \wedge (x \wedge z)))) \vee v \text{ (by modularity since } y \leq y \vee z) \\
&= (x \wedge (y \vee (x \wedge z))) \vee v \text{(since } x \wedge z \leq y \vee z) \\
&= (x \wedge z) \vee (y \wedge z) \vee v \text{ (by modularity since } x \leq x \wedge z) \\
&= v \text{ (due to the definition of } v).
\end{aligned}$$

Similar arguments can be used to prove the remaining equalities. □

**Definition 5.36.** *Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a bounded lattice that has 0 as its least element and 1 as its largest element.*

*The elements $x$ and $y$ are* complementary *if $x \wedge y = 0$ and $x \vee y = 1$.*

If $x$ and $y$ are complementary we say that one element is the *complement* of the other. Lattices in which every element has a complement are referred to as *complemented lattices*.

*Example 5.37.* The lattice $N_5$ is a complemented lattice. Indeed, $x$ and $z$ are complementary elements and so are $y$ and $z$. The lattice $M_5$ is also complemented.

*Example 5.38.* Let $S$ be a set and let $(\mathcal{P}(S), \cap, \cup)$ be the bounded lattice of its subsets having $\emptyset$ as its first element and $S$ as its last element. Unlike the lattices mentioned in Example 5.37, a set $X \in \mathcal{P}(S)$ has a unique complement $S - X$.

*Example 5.39.* Let $(\mathbb{N} \cup \{\infty\}, \leq)$ be the infinite chain of natural numbers extended by $\infty$. If $m, n \in \mathbb{N} \cup \{\infty\}$, then $m \wedge n = \min\{m, n\}$ and $m \vee n = \max\{m, n\}$. Clearly, this is a bounded lattice and no two of elements except 0 and $\infty$ are complementary.

**Theorem 5.40.** *Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a bounded, distributive lattice. For every element $x$, there exists at most one complement.*

*Proof.* Let $x \in L$ and suppose that both $r$ and $s$ are complements of $x$; that is, $x \wedge r = 0$, $x \vee r = 1$, and $x \wedge s = 0$, $x \vee s = 1$. We can write

$$\begin{aligned} r &= r \wedge 1 \\ &= r \wedge (x \vee s) \\ &= (r \wedge x) \vee (r \wedge s) \\ &= 0 \vee (r \wedge s) \\ &= r \wedge s, \end{aligned}$$

which implies $r \leq s$. Similarly, starting with $s$, we obtain

$$\begin{aligned} s &= s \wedge 1 \\ &= s \wedge (r \vee x) \\ &= (s \wedge r) \vee (s \wedge x) \\ &= (s \wedge r) \vee 0 \\ &= s \wedge r, \end{aligned}$$

which implies $s \leq r$. Consequently, $s = r$. □

## 5.4 Complete Lattices

We saw that lattices can be viewed either as algebras or as partially ordered sets. In this section, we focus on a class of lattices that is important for a variety of applications using the second point of view.

**Definition 5.41.** *A* complete lattice *is a poset $(L, \leq)$ such that for every subset $U$ of $L$ both* $\sup U$ *and* $\inf U$ *exist.*

Note that if $U$ and $V$ are two subsets of a complete lattice and $U \subseteq V$, then $\sup U \leq \sup V$ and $\inf V \leq \inf U$. Therefore, for every subset $T$ of $L$, we have

$$\sup \emptyset \leq \sup T \leq \sup L \text{ and } \inf L \leq \inf T \leq \inf \emptyset.$$

If $T$ is a singleton (that is, $T = \{t\}$), then these inequalities amount to

$$\sup \emptyset \leq t \leq \sup L \text{ and } \inf L \leq t \leq \inf \emptyset$$

for every $t \in L$. This means that a complete lattice has a least element $0 = \sup \emptyset = \inf L$ and a greatest element $1 = \inf \emptyset = \sup L$.

For a subset $U$ of the complete lattice, we denote $\sup U$ by $\bigvee U$ and $\inf U$ by $\bigwedge U$.

Obviously, every complete lattice is also a lattice since $x \vee y$ and $x \wedge y$ are $\sup\{x, y\}$ and $\inf\{x, y\}$, respectively.

The associative properties of the usual lattices can be extended to complete lattices as follows. Let $(L, \leq)$ be a complete lattice and let $\mathcal{C} = \{C_i \mid C_i \subseteq L \text{ for } i \in I\}$ be a collection of subsets of $L$. Then,

$$\bigvee_{i \in I} \bigvee C_i = \bigvee \bigcup \mathcal{C},$$
$$\bigwedge_{i \in I} \bigwedge C_i = \bigwedge \bigcup \mathcal{C}.$$

**Theorem 5.42.** *Let $(L, \leq)$ be a poset such that $\sup U$ exists for every subset $U$ of $L$. Then $(L, \leq)$ is a complete lattice.*

*Proof.* It is sufficient to prove that $\inf U$ exists for each subset $U$ of the lattice. By hypothesis, the set $U^i$ of lower bounds of $U$ has a supremum $x = \sup U_i$. Every element of $U$ is an upper bound of $U^i$, which means that $x \leq u$, which implies that $x$ is a lower bound for $U$. Thus, $x \in U^i \cap (U^i)^s$, which means that $x = \inf U$. □

**Theorem 5.43.** *If $(L, \leq)$ is a poset such that $\inf U$ exists for every set $U$, then $(L, \leq)$ is a complete lattice.*

*Proof.* This statement follows by duality from Theorem 5.42. □

*Example 5.44.* Let $S$ be a set. The poset of its subsets $(\mathcal{P}(S), \subseteq)$ is a complete lattice because, for any collection $\mathcal{C}$ of subsets of $S$, $\sup \mathcal{C} = \bigcup \mathcal{C}$ and $\inf \mathcal{C} = \bigcap \mathcal{C}$.

*Example 5.45.* Let $\mathcal{C}$ be a closure system on a set $S$ and let $\mathbf{K}$ be the corresponding closure operator. Then, $(\mathcal{C}, \subseteq)$ is a complete lattice because $\inf \mathcal{D} = \bigcap \mathcal{D}$ and $\sup \mathcal{D} = \mathbf{K}(\bigcup \mathcal{D})$ for any subcollection $\mathcal{D}$ of $\mathcal{C}$.

It is clear that $\inf \mathcal{D}$ exists and equals $\bigcap \mathcal{D}$ for any subcollection $\mathcal{D}$ of $\mathcal{C}$. We show that $\mathbf{K}(\bigcup \mathcal{D})$ equals $\sup \mathcal{D}$. It is clear that $D \subseteq \mathbf{K}(\bigcup \mathcal{D})$. Suppose now that $E$ is a subset of $\mathcal{C}$ that is an upper bound for $\mathcal{D}$, that is, $D \subseteq E$ for every $D \in \mathcal{D}$. We have $\bigcup \mathcal{D} \subseteq E$, so $\mathbf{K}(\bigcup \mathcal{D}) \subseteq \mathbf{K}(E) = E$ because $E \in \mathcal{C}$. Therefore, $\mathbf{K}(\bigcup \mathcal{D})$ is the least upper bound of $\mathcal{D}$.

The notion of a lattice morphism is extended to complete lattices.

**Definition 5.46.** *Let $(L_1, \leq)$ and $(L_2, \leq)$ be two complete lattices. A function $f : L_1 \longrightarrow L_2$ is a* complete lattice morphism *if $f(\bigvee U) = \bigvee f(U)$ and $f(\bigwedge U) = \bigwedge f(U)$ for every subset $U$ of $L_1$.*

*If $f$ is a bijection such that both $f$ and $f^{-1}$ are complete lattice morphisms, then we say that $f$ is a* complete lattice isomorphism.

**Theorem 5.47.** *Every complete lattice is isomorphic to the lattice of closed sets of a closure system.*

*Proof.* Let $(L, \leq)$ be a complete lattice and let $I_x = \{t \in L \mid t \leq x\}$ for $x \in L$. We claim that $\mathfrak{I} = \{I_x \mid x \in L\}$ is a closure system on $L$.

Indeed, note that $I_1$ (where 1 is the largest element of $L$) coincides with $L$, so $L \in \mathfrak{I}$.

Now let $\{I_x \mid x \in M\}$ be an arbitrary family of sets in $\mathfrak{I}$, where $M$ is a subset of $L$. Note that $\bigcap\{I_x \mid x \in M\} = I_y$, where $y = \inf M$. Thus, $\mathfrak{I}$ is a closure system.

It is easy to verify that $f : L \longrightarrow \mathfrak{I}$ given by $f(x) = I_x$ is a complete lattice isomorphism. □

**Definition 5.48.** *Let $(S, \leq)$ and $(T, \leq)$ be two posets. A* Galois connection *between $S$ and $T$ is a pair of mappings $(\phi, \psi)$, where $\phi : S \longrightarrow T$ and $\psi : T \longrightarrow S$ that satisfy the following conditions.*
(i) *If $s_1 \leq s_2$, then $\phi(s_2) \leq \phi(s_1)$ for every $s_1, s_2 \in S$.*
(ii) *If $t_1 \leq t_2$, then $\phi(t_2) \leq \phi(t_1)$ for every $t_1, t_2 \in T$.*
(iii) *$s \leq \psi(\phi(s))$ and $t \leq \phi(\psi(t))$ for $s \in S$ and $t \in T$.*

*Example 5.49.* Let $X$ and $Y$ be two sets and let $\rho$ be a relation, $\rho \subseteq X \times Y$. Define $\phi_\rho : \mathcal{P}(X) \longrightarrow \mathcal{P}(Y)$ and $\psi_\rho : \mathcal{P}(Y) \longrightarrow \mathcal{P}(X)$ by

$$\phi_\rho(U) = \{y \in Y \mid (x, y) \in \rho \text{ for all } x \in U\},$$
$$\psi_\rho(V) = \{x \in X \mid (x, y) \in \rho \text{ for all } y \in V\}$$

for $U \in \mathcal{P}(X)$ and $V \in \mathcal{P}(Y)$.

The pair $(\phi_\rho, \psi_\rho)$ is a Galois connection between the posets $(\mathcal{P}(X), \subseteq)$ and $(\mathcal{P}(Y), \subseteq)$. It is immediate to verify that the first two conditions of Definition 5.48 are satisfied. We discuss here only the third condition of the definition.

To prove that $U \subseteq \psi_\rho(\phi_\rho(U))$, let $u \in U$. We need to show that $(u, y) \in \rho$ for every $y \in \psi_\rho(U)$. By the definition of $\psi_\rho$, if $y \in \psi_\rho(U)$, we have indeed $(u, y) \in \rho$. The proof of the second inclusion of the third part of the definition is similar.

The pair $(\phi_\rho, \psi_\rho)$ will be referred to as the *polarity generated by the relation $\rho$*.

**Theorem 5.50.** *Let $(S, \leq)$ and $(T, \leq)$ be two posets. A pair of mappings $(\phi, \psi)$, where $\phi : S \longrightarrow T$ and $\psi : T \longrightarrow S$, is a Galois connection between $(S, \leq)$ and $(T, \leq)$ if and only if $s \leq \psi(t)$ is equivalent to $t \leq \phi(s)$.*

*Proof.* Suppose that $(\phi, \psi)$ is a pair of mappings such that $s \leq \psi(t)$ is equivalent to $t \leq \phi(s)$. Choosing $t = \phi(s)$, it is clear that $t \leq \phi(s)$, so $s \leq \psi(t) = \psi(\phi(s))$. Similarly, we can show that $t \leq \phi(\psi(t))$, so the pair $(\phi, \psi)$ satisfies the third condition of Definition 5.48.

Let $s_1, s_2 \in S$ such that $s_1 \leq s_2$. Since $s_2 \leq \psi(\phi(s_2))$, we have $s_1 \leq \psi(\phi(s_2))$, which implies $\phi(s_2) \leq \phi(s_1)$. A similar argument can be used to prove that $t_1 \leq t_2$ implies $\psi(t_2) \leq \psi(t_1)$, so $(\phi, \psi)$ satisfies the remaining conditions of the definition, and therefore is a Galois connection.

Conversely, let $(\phi, \psi)$ be a Galois connection. If $s \leq \psi(t)$, then $\phi(\psi(t)) \leq \phi(s)$. Since $t \leq \phi(\psi(t))$, we have $t \leq \phi(s)$. The reverse implication can be shown in a similar manner. □

The notion of a closure operator, which was discussed in Section 4.5, can be generalized to partially ordered sets.

**Definition 5.51.** *Let $(L, \leq)$ be a poset. A mapping $\kappa : L \longrightarrow L$ is a* closure operator *on $L$ if it satisfies the following conditions:*
(i) $u \leq \kappa(u)$ (expansiveness),
(ii) $u \leq v$ *implies* $\kappa(u) \leq \kappa(v)$ (monotonicity), *and*
(iii) $\kappa(\kappa(u)) = \kappa(u)$ (idempotency)
*for $u, v \in L$.*

*Example 5.52.* Let $(S, \leq)$ and $(T, \leq)$ be two posets, and suppose that $(\phi, \psi)$ is a Galois connection between these posets. Then, $\psi\phi$ is a closure on $S$ and $\phi\psi$ is a closure on $T$.

By the third part of Definition 5.48, we have $s \leq \psi(\phi(s))$, so $\psi\phi$ is expansive. Suppose that $s_1 \leq s_2$. This implies $\phi(s_2) \leq \phi(s_1)$, which in turn implies $\psi(\phi(s_1)) \leq \psi(\phi(s_2))$. Thus, $\psi\phi$ is monotonic.

In exactly the same manner, we can prove that $t \leq \phi(\psi(t))$ and that $\phi\psi$ is monotonic.

Since $s \leq \psi(\phi(s))$, we have $\phi(\psi(\phi(s))) \leq \phi(s)$. On the other hand, choosing $t = \phi(s)$, we have $\phi(s) \leq \phi(\psi(\phi(s)))$, so $\phi(s) = \phi(\psi(\phi(s)))$ for every $s \in S$. A similar argument shows that $\psi(t) = \psi(\phi(\psi(t)))$. Therefore we obtain $\psi(\phi(s)) = \psi(\phi(\psi(\phi(s))))$ for every $s \in S$ and $\phi(\psi(t)) = \phi(\psi(\phi(\psi(t))))$, which proves that $\phi\psi$ and $\psi\phi$ are idempotent.

**Lemma 5.53.** *Let $(L, \leq)$ be a complete lattice and let $\kappa : L \longrightarrow L$ be a closure operator. Define the family of $\kappa$-closed elements $Q_\kappa = \{x \in L \mid x = \kappa(x)\}$. Then, $1 \in Q_\kappa$, and for each subset $D$ of $Q_\kappa$, $\bigwedge D \in Q_\kappa$.*

*Proof.* Since $1 \leq \kappa(1) \leq 1$, we have $1 \in Q_\kappa$.

Let $D = \{u_i \mid i \in I\}$ be a collection of elements of $L$ such that $u_i = \kappa(u_i)$ for $i \in I$. Since $\bigwedge D \leq u_i$, we have $\kappa(\bigwedge D) \leq \kappa(u_i) = u_i$ for every $i \in I$. Therefore, $\kappa(\bigwedge D) \leq \bigwedge D$, which implies $\kappa(\bigwedge D) = \bigwedge D$. Thus, $\bigwedge D \in Q_\kappa$. □

**Theorem 5.54.** *Let $(L, \leq)$ be a complete lattice and let $\kappa$ be a closure operator on $L$. Then, $(Q_\kappa, \leq)$ is a complete lattice.*

*Proof.* This statement follows from Lemma 5.53 and from Theorem 5.43. □

If $(\phi, \psi)$ is a Galois connection between the posets $(S, \leq)$ and $(T, \leq)$, then each of the mappings $\phi, \psi$ is an *adjunct* of the other. The next theorem characterizes those mappings between posets that have an adjunct mapping.

**Theorem 5.55.** *Let $(S, \leq)$ and $(T, \leq)$ be two posets and let $\phi : S \longrightarrow T$ be a mapping. There exists a mapping $\psi : T \longrightarrow S$ such that $(\phi, \psi)$ is a Galois connection between $(S, \leq)$ and $(T, \leq)$ if and only if for every $t \in T$ there exists $z \in S$ such that*

$$\phi^{-1}(\{v \in T \mid v \leq t\}) = \{u \in S \mid u \leq z\}.$$

*Proof.* Suppose that the condition of the theorem is satisfied by $\phi$. Given $t \in T$, the element $z \in S$ is unique because the equality $\{u \in S \mid u \leq z\} = \{u \in S \mid u \leq z'\}$ implies $z = z'$. Define the mapping $\psi : T \longrightarrow S$ by $\psi(t) = z$, where $z$ is the element of $S$ whose existence is stipulated by the theorem. Note that $s \leq \psi(t)$ is equivalent to $t \leq \phi(s)$, which means that $(\phi, \psi)$ is a Galois connection according to Theorem 5.50.

The proof of the necessity of the condition of the theorem is immediate. □

## 5.5 Boolean Algebras and Boolean Functions

If $\mathcal{L} = (L, \{\wedge, \vee\})$ is a bounded distributive lattice that is complemented, then, by Theorem 5.40, there is a mapping $h : L \longrightarrow L$ such that $h(x)$ is the complement of $x \in L$. This leads to the following definition.

**Definition 5.56.** *A* Boolean lattice *is a bounded distributive lattice that is complemented.*

An equivalent notion that explicitly introduces two zero-ary operations and one unary operation is the notion of *Boolean algebra*.

**Definition 5.57.** *A* Boolean algebra *is an algebra $\mathcal{B} = (B, \{\wedge, \vee, {}^-, 0, 1\})$ having the type $(2, 2, 1, 0, 0)$ that satisfies the following conditions:*

(i) *$(B, \{\wedge, \vee\})$ is a distributive lattice having $0$ as its least element and $1$ as its greatest element, and*

(ii)$^-$: *$B \longrightarrow B$ is a unary operation such that $\bar{x}$ is the complement of $x$ for $x \in B$.*

Every Boolean algebra has at least two elements, the ones designated by its zero-ary operations.

*Example 5.58.* The two-element Boolean algebra is the Boolean algebra $\mathcal{B}_2 = (\{0, 1\}, \{\wedge, \vee, {}^-, 0, 1\})$ defined by:

$$0 \wedge 0 = 0, 1 \wedge 1 = 1,$$
$$0 \wedge 1 = 1 \wedge 0 = 0,$$
$$0 \vee 0 = 0, 1 \vee 1 = 1,$$
$$0 \vee 1 = 1 \vee 0 = 1,$$
$$\bar{0} = 1, \bar{1} = 0.$$

Example 5.38 can now be recast as introducing a Boolean algebra.

*Example 5.59.* The set $\mathcal{P}(S)$ of subsets of a set $S$ defines a Boolean algebra $(\mathcal{P}(S), \{\cap, \cup, \bar{}, \emptyset, S\})$, where $\overline{X} = S - X$.

*Example 5.60.* Let $\mathcal{B}_4 = (\{0, a, \bar{a}, 1\}, \{\wedge, \vee, \bar{}, 0, 1\})$ be the four-element Boolean algebra whose Hasse diagram is given in Figure 5.4. We leave it to the reader to verify that the poset defined by this diagram is indeed a Boolean algebra.

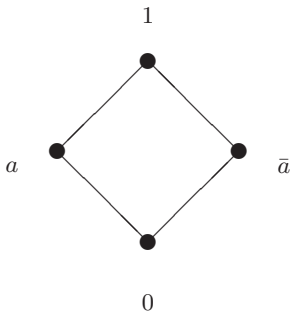



**Fig. 5.4.** Hasse diagram of the four-element Boolean algebra.

The existence of the zero-ary operations means that every subalgebra of a Boolean algebra must contain at least 0 and 1.

In a Boolean algebra $\mathcal{B} = (B, \{\wedge, \vee, \bar{}, 0, 1\})$, we have $\overline{\overline{x}} = x$ because of the symmetry of the definition of the complement and because the complement of an element is unique (since $\mathcal{B}$ is a distributive lattice). This property is known as the *involutive property of the complement.*

**Theorem 5.61 (DeMorgan Laws).** *Let* $\mathcal{B} = (B, \{\wedge, \vee, \bar{}, 0, 1\})$ *be a Boolean algebra. We have*

$$\overline{x \wedge y} = \bar{x} \vee \bar{y}, \overline{x \vee y} = \bar{x} \wedge \bar{y}$$

*for* $x, y \in B$.

*Proof.* By applying the distributivity, commutativity, and associativity of $\wedge$ and $\vee$ operations, we can write

$$\begin{aligned}(\bar{x} \vee \bar{y}) \wedge (x \wedge y) &= (\bar{x} \wedge (x \wedge y)) \vee (\bar{y} \wedge (x \wedge y)) \\ &= ((\bar{x} \wedge x) \wedge y) \vee ((\bar{y} \wedge y) \wedge x) \\ &= (0 \wedge y) \vee (0 \wedge x) = 0\end{aligned}$$

and

$$\begin{aligned}(\bar{x} \vee \bar{y}) \vee (x \wedge y) &= (\bar{x} \vee \bar{y} \vee x) \wedge (\bar{x} \vee \bar{y} \vee y) \\ &= (1 \vee \bar{y}) \wedge (1 \vee \bar{x}) = 1 \vee 1 = 1\end{aligned}$$

for $x, y \in B$. This shows that $\bar{x} \vee \bar{y}$ is the complement of $x \wedge y$; that is, $\overline{x \wedge y} = \bar{x} \vee \bar{y}$.

The second part of the theorem has a similar argument. □

**Definition 5.62.** *Let $\mathcal{B}_i = (B_i, \{\wedge, \vee, \bar{}, 0, 1\})$, $i = 1, 2$, be two Boolean algebras. A* morphism *from $\mathcal{B}_1$ to $\mathcal{B}_2$ is a function $h : B_1 \longrightarrow B_2$ such that*

$$\begin{aligned}h(x \wedge y) &= h(x) \wedge h(y), \\ h(x \vee y) &= h(x) \vee h(y), \\ h(\bar{x}) &= \overline{h(x)},\end{aligned}$$

*for $x, y \in B_1$.*

*An* isomorphism *of Boolean algebras is a morphism that is also a bijection.*

*Example 5.63.* Let $\mathcal{B} = (B, \{\wedge, \vee, \bar{}, 0, 1\})$ be a Boolean algebra and let $c, d \in B$ such that $c \leq d$. We can define a Boolean algebra on the interval $[c, d]$ as

$$\mathcal{B}_{[c,d]} = ([c, d], \{\wedge, \vee, \tilde{}, c, d\}),$$

where $\wedge, \vee$ are the restrictions of the operations of $\mathcal{B}$ to the set $[c, d]$ and $\tilde{x} = (\bar{x} \vee c) \wedge d$ for $x \in B$.

The projection $h : B \longrightarrow [c, d]$ defined by $h(x) = (x \vee c) \wedge d$ for $x \in B$ is a morphism between $\mathcal{B}$ and $\mathcal{B}_{[c,d]}$. We already saw in Example 5.23 that $h$ is a lattice morphism. Thus, we need to prove only that $\overline{h(x)} = h(\bar{x}) = (\bar{x} \vee c) \wedge d$ for $x \in B$. The verification of this equality is left to the reader.

### Boolean Rings

Let $\mathcal{B} = (B, \{\wedge, \vee, \bar{}, 0, 1\})$ be a Boolean algebra and let "$\oplus$" be a binary operation on $B$ defined by

$$a \oplus b = (a \wedge \bar{b}) \vee (\bar{a} \wedge b)$$

for $a, b \in B$.

It is easy to verify that

$$\begin{aligned} a \oplus b &= b \oplus a, \\ (a \oplus b) \oplus c &= a \oplus (b \oplus c), \\ a \oplus a &= 0, \\ a \oplus 1 &= \bar{a}, \\ a \wedge (b \oplus c) &= (a \wedge b) \oplus (a \wedge c), \\ a \wedge 1 &= a, \end{aligned}$$

for every $a, b, c \in B$. Thus, the Boolean algebra $\mathcal{B}$ has a related natural structure of a commutative unitary ring $(B, \{0, 1, \oplus, h, \wedge\})$, where the role of the addition or the ring is played by the operation $\oplus$, the additive inverse is given by $h(a) = a$ for $a \in B$, and each element is idempotent.

*Example 5.64.* For the Boolean algebra $(\mathcal{P}(S), \{\wedge, \vee, {}^{-}, \emptyset, S\})$ introduced in Example 5.59, the additive operation of the ring is the symmetric difference of sets

$$U \oplus V = (U - V) \cup (V - U)$$

for $U, V \in \mathcal{P}(S)$. Thus, we obtain a commutative unitary ring structure $(\mathcal{P}, \{\emptyset, S, \oplus, h, \cap\})$, where $h(U) = U$ for $U \in \mathcal{P}(S)$.

A commutative unitary ring in which each element is its own additive inverse and each element is idempotent defines a Boolean algebra, as we show next.

**Theorem 5.65.** *Let $\mathcal{I} = (B, \{0, 1, +, h, \cdot, 1\}$ be a commutative unitary ring such that $h(b) = b$ and $b \wedge b = b$ for every $b \in B$. Define the operations $\vee, \wedge, {}^{-}$ by*

$$\begin{aligned} a \vee b &= a + b + a \cdot b, \\ a \wedge b &= a \cdot b, \\ \bar{a} &= 1 + a, \end{aligned}$$

*for $a \in B$. Then, $\mathcal{B} = (B, \{\wedge, \vee, {}^{-}, 0, 1\})$ is a Boolean algebra.*

*Proof.* The operation $\vee$ is commutative because $\mathcal{I}$ is a commutative ring. Observe that

$$\begin{aligned} a \vee (b \vee c) &= a \vee (b + c + bc) \\ &= a + b + c + bc + ab + ac + abc, \\ (a \vee b) \vee c &= a + b + ab + c + ac + bc + abc, \end{aligned}$$

which proves that $\vee$ is also associative. Further, we have $a \vee a = a + a + aa = a$, which proves that $\vee$ is idempotent.

The operation "$\wedge$" is known to be commutative, associative, and idempotent since it coincides with the multiplication of the ring $\mathcal{I}$. To prove the distributivity, note that

$$a \wedge (b \vee c) = a(b + c + bc) = ab + ac + abc$$

and

$$(a \wedge b) \vee (a \wedge c) = ab + ac + (ab)(ac) = ab + ac + abc$$

due to the commutativity and idempotency of multiplication in $\mathfrak{I}$. Thus, we have shown that $\mathcal{B} = (B, \{\wedge, \vee, 0, 1\})$ is a distributive lattice having 0 as its least element and 1 as its largest element.

We need to show only that $h(a) = 1 + a$ is the complement of $a$. This is indeed the case because $a \vee (1 + a) = a + 1 + a + a(1 + a) = 1$ and $a \wedge h(a) = a(1 + a) = a + a = 0$ for every $a \in B$. □

### Boolean Functions

**Definition 5.66.** *Let $\mathcal{B} = (B, \{\wedge, \vee, \bar{}, 0, 1\})$ be a Boolean algebra. For $n \in \mathbb{N}$, the set $BF(\mathcal{B}, n)$ of* Boolean functions *of $n$ arguments over $\mathcal{B}$ contains the following functions:*

(i) *For every $b \in B$, the constant function $f_b : B^n \longrightarrow B$ defined by*

$$f_b(x_0, \ldots, x_{n-1}) = b$$

*for every $x_0, \ldots, x_{n-1} \in B$ belongs to $BF(\mathcal{B}, n)$.*

(ii) *Every* projection function *$p_i^n : B^n \longrightarrow B$ given by $p_i^n(x_0, \ldots, x_{n-1}) = x_i$ for every $x_0, \ldots, x_{n-1} \in B$ belongs to $BF(\mathcal{B}, n)$.*

(iii) *If $f, g \in BF(\mathcal{B}, n)$, then the functions $f \wedge g, f \vee g$, and $\bar{f}$ given by*

$$(f \wedge g)(x_0, \ldots, x_{n-1}) = f(x_0, \ldots, x_{n-1}) \wedge g(x_0, \ldots, x_{n-1}),$$
$$(f \vee g)(x_0, \ldots, x_{n-1}) = f(x_0, \ldots, x_{n-1}) \vee g(x_0, \ldots, x_{n-1}),$$

*and*

$$\bar{f}(x_0, \ldots, x_{n-1}) = \overline{f(x_0, \ldots, x_{n-1})}$$

*for every $x_0, \ldots, x_{n-1} \in B$ belong to $BF(\mathcal{B}, n)$.*

**Definition 5.67.** *For $n \in \mathbb{N}$, the set $SBF(\mathcal{B}, n)$ of* simple Boolean functions *of $n$ arguments consists of the following functions:*

(i) *Every projection function $p_i^n : B^n \longrightarrow B$ given by $p_i^n(x_0, \ldots, x_{n-1}) = x_i$ for every $x_0, \ldots, x_{n-1} \in B$.*

(ii) *If $f, g \in SBF(\mathcal{B}, n)$, then the functions $f \wedge g, f \vee g$ and $\bar{f}$ given by*

$$(f \wedge g)(x_0, \ldots, x_{n-1}) = f(x_0, \ldots, x_{n-1}) \wedge g(x_0, \ldots, x_{n-1}),$$
$$(f \vee g)(x_0, \ldots, x_{n-1}) = f(x_0, \ldots, x_{n-1}) \vee g(x_0, \ldots, x_{n-1}),$$

*and*

$$\bar{f}(x_0, \ldots, x_{n-1}) = \overline{f(x_0, \ldots, x_{n-1})}$$

*for every $x_0, \ldots, x_{n-1} \in B$ belong to $SBF(\mathcal{B}, n)$.*

If $x \in B$ and $a \in \{0,1\}$, define the function $x^a$ as

$$x^a = \begin{cases} x & \text{if } a = 1 \\ \bar{x} & \text{if } a = 0. \end{cases}$$

Observe that $x = a$ if and only if $x^a = 1$, and $x = \bar{a}$ if and only if $x^a = 0$.

For $\mathbf{A} = (a_1, \ldots, a_n) \in \{0,1,*\}^n$, define the simple Boolean function $t^{\mathbf{A}} : B^n \longrightarrow B$ as

$$t^{\mathbf{A}}(x_1, \ldots, x_n) = x_{i_1}^{a_{i_1}} \wedge x_{i_2}^{a_{i_2}} \wedge x_{i_p}^{a_{i_p}}$$

for $(x_1, \ldots, x_n) \in B^n$, where $\{a_{i_1}, \ldots, a_{i_p}\} = \{a_i \mid a_i \neq *, 1 \leq i \leq n\}$. This function is an *n-ary term* for the Boolean algebra $\mathcal{B}$. The set of $n$-ary terms of a Boolean algebra is denoted by $\mathtt{T}(\mathcal{B}, n)$.

Those components of $\mathbf{A}$ that equal $*$ denote the places of variables that do not appear in the term $t^{\mathbf{A}}$.

*Example 5.68.* Let $\mathbf{A} = (1, *, 0, 0, *) \in \{0,1\}^n$. The 5-term $t^{\mathbf{A}}$ is

$$t(x_1, x_2, x_3, x_4, x_5) = x_1 \wedge \bar{x}_3 \wedge \bar{x}_4$$

for every $(x_1, x_2, x_3, x_4, x_5) \in B^5$.

It is easy to see that if $\mathbf{A}, \mathbf{B} \in \{0,1\}^n$, then

$$t^{\mathbf{B}}(\mathbf{A}) = \begin{cases} 1 & \text{if } \mathbf{A} = \mathbf{B}, \\ 0 & \text{if } \mathbf{A} \neq \mathbf{B}. \end{cases}$$

To simplify the notation, whenever there is no risk of confusion, we omit the symbol "$\wedge$" and denote an application of this operation by a simple juxtaposition of symbols. For example, instead of writing $a \wedge b$, we use the notation $ab$. For the same reason, we will assume thar $\wedge$ has higher priority than $\vee$. These assumptions allow us to write $a \vee bc$ instead of $a \vee (b \wedge c)$.

**Theorem 5.69.** *Let $\mathcal{B} = (B, \{\wedge, \vee, \bar{}, 0, 1\})$ be a Boolean algebra. For every $(x_1, \ldots, x_n) \in B^n$, where $n \geq 1$, we have*
(i) $t^{\boldsymbol{A}}(x_1, \ldots, x_n) t^{\boldsymbol{B}}(x_1, \ldots, x_n) = 0$ *for* $\boldsymbol{A}, \boldsymbol{B} \in \{0,1\}^n$ *and* $\boldsymbol{A} \neq \boldsymbol{B}$,
(ii) $\bigvee\{t^{\boldsymbol{A}}(x_1, \ldots, x_n) \mid \boldsymbol{A} \in \{0,1\}^n\} = 1$, *and*
(iii) $\overline{t^{\boldsymbol{A}}(x_1, \ldots, x_n)} = \bigvee\{t^{\boldsymbol{B}}(x_1, \ldots, x_n) \mid \boldsymbol{B} \in \{0,1\}^n - \{\boldsymbol{A}\}\}$
*for every $(x_1, \ldots, x_n) \in B^n$.*

*Proof.* Let $\mathbf{A} = (a_0, \ldots, a_{n-1})$ and $\mathbf{B} = (b_0, \ldots, b_{n-1})$. If $\mathbf{A} \neq \mathbf{B}$, then there exists $i$ such that $0 \leq i \leq n-1$ and $a_i \neq b_i$. Therefore, by applying the commutativity and associativity properties of $\wedge$, the expression

$$(t^{\mathbf{A}} t^{\mathbf{B}})(x_1, \ldots, x_n) = x_1^{a_1} \wedge \cdots \wedge x_n^{a_n} \wedge x_1^{b_1} \wedge \cdots \wedge x_n^{b_n}$$

can be written as

$$(t^{\mathbf{A}}t^{\mathbf{B}})(x_1,\ldots,x_n) = \cdots \wedge x_i^{a_i} \wedge x_i^{b_i} \wedge \cdots = \cdots \wedge x_i^1 \wedge x_i^0 \wedge \cdots = 0.$$

The proof of the second part can be done by induction on $n$. In the base case, $n = 1$, the desired equality amounts to $x_1^0 \vee x_1^1 = 1$, which obviously holds.

Suppose now that the equality holds for $n$. We have

$$\begin{aligned}
&\bigvee\{(t^{\mathbf{A}}(x_1,\ldots,x_{n+1}) \mid \mathbf{A} \in \{0,1\}^{n+1}\}\\
&= \bigvee\{(x_1,\ldots,x_n)^{(a_1,\ldots,a_n)} \wedge x_{n+1}^0 \mid (a_1,\ldots,a_n) \in \{0,1\}^n\}\\
&\quad \bigvee\{(x_1,\ldots,x_n)^{(a_1,\ldots,a_n)} \wedge x_{n+1}^1 \mid (a_1,\ldots,a_n) \in \{0,1\}^n\}\\
&= \bigvee\{(x_1,\ldots,x_n)^{(a_1,\ldots,a_n)} \mid (a_1,\ldots,a_n) \in \{0,1\}^n\} \wedge (x_{n+1}^0 \vee x_{n+1}^1)\\
&= \bigvee\{(x_1,\ldots,x_n)^{(a_1,\ldots,a_n)} \mid (a_1,\ldots,a_n) \in \{0,1\}^n\}\\
&= 1 \text{ (by the inductive hypothesis).}
\end{aligned}$$

Part (iii) of the theorem follows by observing that

$$\overline{t^{\mathbf{A}}(x_1,\ldots,x_n)} = \begin{cases} 1 & \text{if } (x_1,\ldots,x_n) \neq \mathbf{A},\\ 0 & \text{if } (x_1,\ldots,x_n) = \mathbf{A}. \end{cases}$$

The right-hand member of the equality takes exactly the same values, as can be seen easily. □

The set $BF(\mathcal{B}, n)$ is itself a Boolean algebra relative to the operations $\vee$, $\wedge$, and $^-$ from Definition 5.66. The least element is the constant function $f_0 : B^n \longrightarrow B$ given by $f_0(x_1,\ldots,x_n) = 0$, and the largest element is the constant function $f_1 : B^n \longrightarrow B$ given by $f_1(x_1,\ldots,x_n) = 0$ for $(x_1,\ldots,x_n) \in B^n$.

A partial order on $BF(\mathcal{B}, n)$ can be introduced by defining $f \leq g$ if $f(x_1,\ldots,x_n) \leq g(x_1,\ldots,x_n)$ for $x_1,\ldots,x_n \in B$. It is clear that $f \leq g$ if and only if $f \vee g = g$ or $f \wedge g = f$.

**Theorem 5.70.** *Let $\mathcal{B} = (B, \{\wedge, \vee, ^-, 0, 1\})$ be a Boolean algebra, $\boldsymbol{A}, \boldsymbol{B} \in \{0, 1, *\}^n$, and let $t^{\boldsymbol{A}}$ and $t^{\boldsymbol{B}}$ be the terms in $\mathrm{T}(\mathcal{B}, n)$ that correspond to $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. We have $t^{\boldsymbol{A}} \leq t^{\boldsymbol{B}}$ if and only if $a_k = *$ implies $b_k = *$ and $a_k \neq *$ implies $a_k = b_k$ or $b_k = *$ for $1 \leq k \leq n$.*

*Proof.* Suppose that $t^{\mathbf{A}} \leq t^{\mathbf{B}}$; that is,

$$x_{i_1}^{a_{i_1}} \wedge x_{i_2}^{a_{i_2}} \wedge x_{i_p}^{a_{i_p}} \leq x_{j_1}^{b_{j_1}} \wedge x_{j_2}^{b_{j_2}} \wedge x_{j_q}^{b_{j_q}},$$

for $(x_1,\ldots,x_n) \in B^n$. Here $\{i_1,\ldots,i_p\} = \{i \mid 1 \leq i \leq n \text{ and } a_i \neq *\}$ and $\{j_1,\ldots,j_q\} = \{j \mid 1 \leq j \leq n \text{ and } b_j \neq *\}$.

Suppose that $a_k = *$ but $b_k \in \{0, 1\}$. Choose $x_{i_\ell} = a_{i_\ell}$ for $1 \leq \ell \leq p$ and $x_k = \bar{b}_k$. The remaining components of $(x_1,\ldots,x_k)$ can be chosen arbitrarily. Clearly, $t^{\mathbf{A}}(x_1,\ldots,x_n) = 1$ and $t^{\mathbf{B}}(x_1,\ldots,x_n) = 0$ because $x_k^{b_k} = 0$. This contradicts the inequality $t^{\mathbf{A}} \leq t^{\mathbf{B}}$, so we must have $b_k = *$.

Suppose now that $a_k \in \{0,1\}$ and $b_k \neq *$. This means that $x_k$ occurs in both $t^{\mathbf{A}}$ and $t^{\mathbf{B}}$, so there exists $i_r = j_s = k$ for some $r, s$, $1 \leq r \leq p$ and $1 \leq s \leq q$. Choose as before $x_{i_\ell} = a_{i_\ell}$ for $1 \leq \ell \leq p$, which implies $t^{\mathbf{A}}(x_1, \ldots, x_n) = 1$, which in turn implies $t^{\mathbf{B}}(x_1, \ldots, x_n) = 1$. This is possible only if $b_k = a_k$, which concludes the argument. □

**Corollary 5.71.** *The minimal elements of the poset* $(\mathtt{T}(\mathcal{B}, n), \leq)$ *are terms that depend on all their arguments, that is, terms of the form*

$$t^{\boldsymbol{B}}(x_1, \ldots, x_n) = x_1^{b_1} x_2^{b_2} \cdots x_n^{b_n}$$

*for* $(x_1, \ldots, x_n) \in B^n$.

*Proof.* If $t^{\mathbf{B}}$ is a minimal element of $(\mathtt{T}(\mathcal{B}, n), \leq)$, then $t^{\mathbf{A}} \leq t^{\mathbf{B}}$ implies $t^{\mathbf{A}} = t^{\mathbf{B}}$. Suppose that there is $k$ such that $b_k = *$. Then, by defining

$$a_i = \begin{cases} b_i & \text{if } i \neq k, \\ 0 \text{ or } 1 & \text{otherwise,} \end{cases}$$

we would have $t^{\mathbf{A}} < t^{\mathbf{B}}$, which would contradict the minimality of $t^{\mathbf{B}}$. □

The minimal terms of the poset $(\mathtt{T}(\mathcal{B}, n), \leq)$, described by Corollary 5.71 are known as $n$-ary *minterms*.

**Definition 5.72.** *Let* $\mathcal{B} = (B, \{\wedge, \vee, \bar{\ }, 0, 1\})$ *be a Boolean algebra and let* $f : B^n \longrightarrow B$ *be a Boolean function. A* disjunctive normal form *of* $f$ *is an expression of the form* $\bigvee_{i=1}^{k} t^{\boldsymbol{A}_i}(x_1, \ldots, x_n) b_{\boldsymbol{A}_i}$, *where* $\boldsymbol{A}_i \in \{0, 1, *\}^n$, $\{b_{\boldsymbol{A}_1}, \ldots, b_{\boldsymbol{A}_k}\} \subseteq B$, *and*

$$f(x_1, \ldots, x_n) = \bigvee_{i=1}^{k} t^{\boldsymbol{A}_i}(x_1, \ldots, x_n) b_{\boldsymbol{A}_i}$$

*for* $(x_1, \ldots, x_n) \in B^n$.

We can prove the existence of a special disjunctive normal form for every Boolean function, which involves only minterms.

**Theorem 5.73.** *Let* $\mathcal{B} = (B, \{\wedge, \vee, \bar{\ }, 0, 1\})$ *be a Boolean algebra. A function* $f : B^n \longrightarrow B$ *is a Boolean function if and only if there exists a family* $\{b_{\boldsymbol{A}} \mid \boldsymbol{A} \in \{0,1\}^n\}$ *of elements of* $B$

$$f(x_1, \ldots, x_n) = \bigvee_{\boldsymbol{A} \in \{0,1\}^n} t^{\boldsymbol{A}}(x_1, \ldots, x_n) b_{\boldsymbol{A}} \tag{5.10}$$

*for every* $(x_1, \ldots, x_n) \in B^n$.

*Proof.* The sufficiency of this condition is obvious. The necessity will be shown by induction on the definition of Boolean functions.

For the base case, we need to consider constant functions and projections. Let $\mathbf{A} = (a_0, \ldots, a_{n-1}) \in \{0,1\}^n$. For a constant function $f_a(x_1, \ldots, x_n) = a$ for $(x_1, \ldots, x_n)$, we can define $b_{\mathbf{A}} = a$ for every $\mathbf{A} \in \{0,1\}^n$ because

$$f(x_1, \ldots, x_n) = a = \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n) a$$

by the second part of Theorem 5.69.

For a projection $p_i^n : B^n \longrightarrow B$ given by $p_i^n(x_1, \ldots, x_n) = x_i$ for $(x_1, \ldots, x_n) \in B^n$, let $b_{\mathbf{A}}$ be

$$b_{\mathbf{A}} = \begin{cases} 1 & \text{if } a_i = 1, \\ 0 & \text{otherwise,} \end{cases}$$

for $\mathbf{A} \in \{0,1\}^n$. We have

$$\begin{aligned} &\bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n) b_{\mathbf{A}} \\ &= \bigvee_{\mathbf{A} \in \{0,1\}^n} t_{(a_1, \ldots, a_{i-1}, 1, a_{i+1}, \ldots, a_n)}(x_1, \ldots, x_n) \\ &= x_i \bigvee_{\mathbf{A} \in \{0,1\}^{n-1}} (x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{n-1})^{(a_0, \ldots, a_{i-1}, a_{i+1}, \ldots, a_{n-1})} \\ &= x_i = p_i^n(x_1, \ldots, x_n). \end{aligned}$$

For the inductive step, suppose that the statement holds for the functions $f, g \in BF(\mathcal{B}, n)$; that is,

$$\begin{aligned} f(x_1, \ldots, x_n) &= \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n) b_{\mathbf{A}}, \\ g(x_1, \ldots, x_n) &= \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n) c_{\mathbf{A}}, \end{aligned}$$

for $(x_1, \ldots, x_n) \in B^n$. Then, $f \vee g$ is

$$(f \vee g)(x_1, \ldots, x_n) = \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n)(b_{\mathbf{A}} \vee c_{\mathbf{A}})$$

by the associativity, commutativity, and idempotency of "$\vee$".

For $f \wedge g$, we can write

$$(f \wedge g)(x_1, \ldots, x_n) = \left( \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n) b_{\mathbf{A}} \right) \wedge \left( \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n) c_{\mathbf{A}} \right) = \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n)(b_{\mathbf{A}} \wedge c_{\mathbf{A}})$$

by applying the distributivity properties of the operations $\vee$ and $\wedge$.

For $\overline{f}$, we have

$$\begin{aligned} \overline{f(x_1, \ldots, x_n)} &= \bigwedge_{\mathbf{A} \in \{0,1\}^n} \left( \overline{t^{\mathbf{A}}(x_1, \ldots, x_n)} \vee \overline{b_{\mathbf{A}}} \right) \\ &= \bigwedge_{\mathbf{A} \in \{0,1\}^n} \left( \bigvee \{ t^{\mathbf{B}}(x_1, \ldots, x_n) \mid \mathbf{B} \in \{0,1\}^n - \{\mathbf{A}\} \} \vee \overline{b_{\mathbf{A}}} \right) \\ &= \bigwedge_{\mathbf{A} \in \{0,1\}^n} \left( \bigvee \{ t^{\mathbf{B}}(x_1, \ldots, x_n) \mid \mathbf{B} \in \{0,1\}^n - \{\mathbf{A}\} \} \vee \overline{b_{\mathbf{A}}} \right). \end{aligned}$$

For $\mathbf{C}, \mathbf{D} \in \{0,1\}^n$ and $(x_1, \ldots, x_n) \in B^n$, define $\phi_{\mathbf{C},\mathbf{D}}(x_1, \ldots, x_n)$ as

$$\phi_{\mathbf{C},\mathbf{D}}(x_1, \ldots, x_n) = \begin{cases} t^{\mathbf{D}}(x_1, \ldots, x_n) & \text{if } \mathbf{D} \neq \mathbf{C}, \\ \overline{b_{\mathbf{C}}} & \text{if } \mathbf{D} = \mathbf{C}. \end{cases}$$

Then, we can write

$$\begin{aligned} \overline{f(x_1, \ldots, x_n)} &= \bigwedge_{\mathbf{A} \in \{0,1\}^n} \bigvee_{\mathbf{D} \in \{0,1\}^n} \phi_{\mathbf{A},\mathbf{D}}(x_1, \ldots, x_n) \\ &= \bigvee_{\mathbf{D} \in \{0,1\}^n} \bigwedge_{\mathbf{A} \in \{0,1\}^n} \phi_{\mathbf{A},\mathbf{D}}(x_1, \ldots, x_n) \\ &\quad \text{(by the distributivity property)} \\ &= \bigvee_{\mathbf{D} \in \{0,1\}^n} (t^{\mathbf{D}}(x_1, \ldots, x_n) \wedge \overline{b_{\mathbf{D}}}), \end{aligned}$$

which concludes the argument. □

Equality (5.10) is known as the *standard disjunctive normal form* of the Boolean function $f$.

Note that by replacing $(x_1, \ldots, x_n)$ by $\mathbf{C} = (c_1, \ldots, c_n) \in \{0,1\}^n$ in Equality (5.10), we obtain $f(c_1, \ldots, c_n) = b_{\mathbf{C}}$, which shows that the elements of the form $b_{\mathbf{A}}$, known as the *standard disjunctive coefficients*, are uniquely determined by the function $f$. Now, we can rewrite Equality (5.10) as

$$f(x_1, \ldots, x_n) = \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n) f(\mathbf{A})$$

for every $(x_1, \ldots, x_n) \in B^n$.

Consider the standard disjunctive normal form of the Boolean function $\overline{f} : B^n \longrightarrow B$ given by

$$\overline{f}(x_1, \ldots, x_n) = \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n)\overline{f(\mathbf{A})}.$$

By applying the $^-$ operation in both members, we can write

$$\begin{aligned} f(x_1, \ldots, x_n) &= \bigwedge_{\mathbf{A} \in \{0,1\}^n} (\overline{t^{\mathbf{A}}}(x_1, \ldots, x_n) \vee f(\mathbf{A})) \\ &= \bigwedge_{(a_1, \ldots, a_n) \in \{0,1\}^n} \left( \bigvee_{i=1}^{n} x_i^{\bar{a}_i} \vee f(a_1, \ldots, a_n) \right) \\ &= \bigwedge_{(\bar{a}_1, \ldots, \bar{a}_n) \in \{0,1\}^n} \left( \bigvee_{i=1}^{n} x_i^{a_i} \vee f(\bar{a}_1, \ldots, \bar{a}_n) \right) \\ &= \bigwedge_{(a_1, \ldots, a_n) \in \{0,1\}^n} \left( \bigvee_{i=1}^{n} x_i^{a_i} \vee f(\bar{a}_1, \ldots, \bar{a}_n) \right). \end{aligned}$$

The last equality is known as the *conjunctive normal form* of the function $f$.

The existence of the standard disjunctive normal form shows that a Boolean function $f : B^n \longrightarrow B$ is completely determined by its values on $n$-tuples $\mathbf{A} \in \{0,1\}^n$. Thus, to fully specify a Boolean function, we can use a table that has $2^n$ rows, one for each $n$-tuple $\mathbf{A}$.

*Example 5.74.* Consider the Boolean function $f : B^3 \longrightarrow B$ given by

| $x_1$ | $x_2$ | $x_3$ | $f(x_1, x_2, x_3)$ |
|---|---|---|---|
| 0 | 0 | 0 | $a$ |
| 0 | 0 | 1 | $a$ |
| 0 | 1 | 0 | $a$ |
| 0 | 1 | 1 | $b$ |
| 1 | 0 | 0 | $a$ |
| 1 | 0 | 1 | $b$ |
| 1 | 1 | 0 | $b$ |
| 1 | 1 | 1 | $b$ |

Its standard disjunctive normal form is

$$\begin{aligned} f(x_1, x_2, x_3) = {} & t^{(0,0,0)}(x_1, x_2, x_3)a \vee t^{(0,0,1)}(x_1, x_2, x_3)a \vee t^{(0,1,0)}(x_1, x_2, x_3)a \\ & \vee t^{(0,1,1)}(x_1, x_2, x_3)b \vee t^{(1,0,0)}(x_1, x_2, x_3)a \vee t^{(1,0,1)}(x_1, x_2, x_3)b \\ & \vee t^{(1,1,0)}(x_1, x_2, x_3)b \vee t^{(1,1,1)}(x_1, x_2, x_3)b. \end{aligned}$$

**Theorem 5.75.** *Let $\mathcal{B} = (B, \{\wedge, \vee, ^-, 0, 1\})$ be a Boolean algebra and let $f, g \in BF(\mathcal{B}, n)$. We have $f \leq g$ if and only if $f(\boldsymbol{A}) \leq g(\boldsymbol{A})$ for every $\boldsymbol{A} \in \{0,1\}^n$.*

*Proof.* The necessity of the condition is obvious. Suppose that $f(\mathbf{A}) \leq g(\mathbf{A})$ for every $\mathbf{A} \in \{0,1\}^n$. Then, by the monotonicity of the binary operations of the Boolean algebra, we have

$$\begin{aligned} f(x_1,\ldots,x_n) &= \bigvee_{\mathbf{A}\in\{0,1\}^n} t^{\mathbf{A}}(x_1,\ldots,x_n) f(\mathbf{A}) \\ &\leq \bigvee_{\mathbf{A}\in\{0,1\}^n} t^{\mathbf{A}}(x_1,\ldots,x_n) g(\mathbf{A}) = g(x_1,\ldots,x_n) \end{aligned}$$

for $x_1,\ldots,x_n \in B^n$, which gives the desired inequality. □

The next theorem (see [114]) is a characterization of simple Boolean functions.

**Theorem 5.76.** *Let $\mathcal{B} = (B, \{\wedge, \vee, \bar{}, 0, 1\})$ be a Boolean algebra. The following statements that concern a function $f : B^n \longrightarrow B$ are equivalent:*
(i) *$f$ is a simple Boolean function.*
(ii) *$f$ is a Boolean function, and $f(\boldsymbol{A}) \in \{0,1\}$ for every $\boldsymbol{A} \in \{0,1\}^n$.*
(iii) *$f(x_1,\ldots,x_n) = 0$ for every $(x_1,\ldots,x_n) \in B^n$ or*

$$f(x_1,\ldots,x_n) = \bigvee\{(x_1,\ldots,x_n)^{\boldsymbol{A}} \mid f(\boldsymbol{A}) = 1\}.$$

*Proof.* (i) implies (ii): Clearly every simple Boolean function is a Boolean function. The proof that $f(\mathbf{A}) \in \{0,1\}$ for every $\mathbf{A} \in \{0,1\}^n$ is by induction on the definition of simple Boolean functions and is left to the reader.

(ii) implies (iii): This implication follows from the existence of the standard disjunctive normal form of Boolean functions.

(iii) implies (i): The constant function $f_0(x_1,\ldots,x_n) = 0$ for $(x_1,\ldots,x_n) \in B^n$ can be written as $f_0(x_1,\ldots,x_n) = x_1 \wedge \bar{x_1}$, so $f_0$ is a simple Boolean function. It is clear that if $f(x_1,\ldots,x_n) = \bigvee\{(x_1,\ldots,x_n)^{\mathbf{A}} \mid f(\mathbf{A}) = 1\}$, then $f$ is a simple Boolean function. □

For a Boolean algebra $\mathcal{B} = (B, \{\wedge, \vee, \bar{}, 0, 1\})$ with $|B| = k$ there exist $k^{k^n}$ functions of the form $f : B^n \longrightarrow B$. The number of Boolean functions can be considerably smaller. Indeed, since a Boolean function $f$ is completely determined by the collection $\{f(\mathbf{A}) \mid A \in \{0,1\}^n\}$, it follows that the number of Boolean functions in $BF(\mathcal{B}, n)$ is $2^{2^n}$. For example, if $\mathcal{B}$ is the four-element Boolean algebra from Example 5.60, there are $4^{4^5} = 2^{2048}$ functions of five arguments defined on the Boolean algebra. However, only $2^{32}$ of these functions are Boolean functions.

### Binary Boolean Functions

**Definition 5.77.** *Let $\mathcal{B}_2 = (\{0,1\}, \{\wedge, \vee, \bar{}, 0, 1\})$ be the two-element Boolean algebra. A* binary Boolean function *is a function $f : \{0,1\}^n \longrightarrow \{0,1\}$.*

We saw that in general Boolean algebra there are many functions that are not Boolean. However, in two-element Boolean algebras, any function is a Boolean function, as we show next.

**Theorem 5.78.** *Every function $f : \{0,1\}^n \longrightarrow \{0,1\}$ is a binary Boolean function in the two-element Boolean algebra $\mathcal{B}_2$.*

*Proof.* Consider the binary Boolean function $g : \{0,1\}^n \longrightarrow \{0,1\}$ defined by $g(x_1, \ldots, x_n) = \bigvee_{\mathbf{A} \in \{0,1\}^n} (x_1, \ldots, x_n)^{\mathbf{A}} f(\mathbf{A})$. It is clear that $g(\mathbf{A}) = f(\mathbf{A})$ for every $\mathbf{A} \in \{0,1\}^n$, so $g = f$. Thus, $f = \bigvee_{\mathbf{A} \in \{0,1\}^n} (x_1, \ldots, x_n)^{\mathbf{A}} f(\mathbf{A})$, which implies that $f$ is indeed a Boolean function. □

**Definition 5.79.** *An* implicant *of a binary Boolean function $f : \{0,1\}^n \longrightarrow \{0,1\}$ is a term $t^{\boldsymbol{A}} \in \mathtt{T}(\mathcal{B}_2, n)$ such that $t^{\boldsymbol{A}} \leq f$.*

*The* rank of an implicant $t^{\boldsymbol{A}}$ *of $f : \{0,1\}^n \longrightarrow \{0,1\}$ is the number $r(t^{\boldsymbol{A}}) = |\{i \mid 1 \leq i \leq n, a_i = *\}|$. Observe that implicants with higher ranks contain fewer literals than implicants with lower rank.*

*The set of implicants of rank $k$ of $f$, $0 \leq k \leq n$, is the set $L_f^k$ that consists of all implicants of rank $k$ for $f$.*

*The set of implicants of $f$ will be denoted by $IMPL_f$. For $f : \{0,1\}^n \longrightarrow \{0,1\}$, we have $IMPL_f = \bigcup_{k=0}^{n-1} L_f^k$.*

Starting from the standard disjunctive normal form for a function $f : B^n \longrightarrow B$,

$$f(x_1, \ldots, x_n) = \bigvee_{\mathbf{A} \in \{0,1\}^n} t^{\mathbf{A}}(x_1, \ldots, x_n) f(\mathbf{A}),$$

it follows that if $f(\mathbf{A}) = 1$, then the minterm $t^{\mathbf{A}}$ is an implicant of $f$ in $L_f^0$. Furthermore, each such term is a minimal implicant of $f$ (relative to the partial order introduced on $\mathtt{T}(\mathcal{B}_2, n)$).

In the next definition, we introduce a partial operation on the set $\mathtt{T}(\mathcal{B}_2, n)$.

**Definition 5.80.** *Let $\boldsymbol{A}, \boldsymbol{B} \in \{0,1,*\}^n$ be two $n$-tuples. Suppose that there exists $k$, $1 \leq k \leq n$ such that*

1. *$a_i = b_i$ if $1 \leq i \leq n$ and $i \neq k$;*
2. *$a_k, b_k \in \{0,1\}$ and $a_k = \overline{b}_k$.*

*The* consensus *of the terms $t^{\boldsymbol{A}}$ and $t^{\boldsymbol{B}}$ is the term $t^{\boldsymbol{C}}$, where $\boldsymbol{C} = (c_1, \ldots, c_n)$ and*

$$c_i = \begin{cases} a_i = b_i & \text{if } i \neq k, \\ * & \text{otherwise,} \end{cases}$$

*for $1 \leq i \leq n$.*

*The consensus of $t^{\boldsymbol{A}}$ and $t^{\boldsymbol{C}}$ is denoted by $t^{\boldsymbol{A}} \boxplus t^{\boldsymbol{B}}$.*

Observe that if the consensus $t^{\mathbf{C}}$ of the terms $t^{\mathbf{A}}$ and $t^{\mathbf{A}}$ exists, then $r(t^{\mathbf{C}}) = r(t^{\mathbf{A}}) + 1 = r(t^{\mathbf{B}}) + 1$. Furthermore, it is immediate that $t^{\mathbf{C}} = t^{\mathbf{A}} \vee t^{\mathbf{B}}$ in the Boolean algebra of Boolean functions.

*Example 5.81.* Let $t^{\mathbf{A}}$ and $t^{\mathbf{B}}$ be the terms

$$t^{\mathbf{A}} = x_1 \wedge \bar{x}_3 \wedge \bar{x}_4 \wedge x_6,$$
$$t^{\mathbf{B}} = x_1 \wedge x_3 \wedge \bar{x}_4 \wedge x_6,$$

from $\mathtt{T}(\mathcal{B}_2, 6)$. Their consensus is the term

$$\begin{aligned} &t^{\mathbf{A}}(x_1, \ldots, x_6) \vee t^{\mathbf{B}}(x_1, \ldots, x_6) \\ &= (x_1 \wedge \bar{x}_3 \wedge \bar{x}_4 \wedge x_6) \vee (x_1 \wedge x_3 \wedge \bar{x}_4 \wedge x_6) \\ &= x_1 \wedge \bar{x}_4 \wedge x_6. \end{aligned}$$

**Theorem 5.82.** *Let $f : \{0,1\}^n \longrightarrow \{0,1\}$ be a Boolean function. If $t^{\mathbf{A}}$ and $t^{\mathbf{B}}$ are implicants of $f$ and their consensus $t^{\mathbf{C}} = t^{\mathbf{A}} \vee t^{\mathbf{B}}$ exists, then $t^{\mathbf{C}}$ is also an implicant of $f$.*

*Proof.* The existence of the consensus $t^{\mathbf{C}}$ of $t^{\mathbf{A}}$ and $t^{\mathbf{B}}$ means that there exists $k$, $1 \leq k \leq n$ such that $a_i = b_i$ if $1 \leq i \leq n$ and $a \neq k$, $a_k, b_k \in \{0,1\}$, and $a_k = \bar{b}_k$.

Since both $t^{\mathbf{A}}$ and $t^{\mathbf{B}}$ are implicants of $f$, it follows that $t^{\mathbf{A}}(x_1, \ldots, x_n) \leq f(x_1, \ldots, x_n)$ and $t^{\mathbf{B}}(x_1, \ldots, x_n) \leq f(x_1, \ldots, x_n)$ for every $(x_1, \ldots, x_n) \in \{0,1\}^n$. Thus,

$$t^{\mathbf{C}}(x_1, \ldots, x_n) = t^{\mathbf{A}}(x_1, \ldots, x_n) \vee t^{\mathbf{B}}(x_1, \ldots, x_n) \leq f(x_1, \ldots, x_n),$$

which means that $t^{\mathbf{C}}$ is an implicant of $f$. □

**Definition 5.83.** *A* prime implicant *of a function $f : \{0,1\}^n \longrightarrow \{0,1\}$ is a maximal element of the poset $(IMPL_f, \leq)$.*

**Theorem 5.84.** *For every binary Boolean function $f : \{0,1\}^n \longrightarrow \{0,1\}$, we have*

$$L_f^{k+1}(\varphi) = \{t^{\mathbf{A}} \vee t^{\mathbf{B}} \mid t^{\mathbf{A}} \vee t^{\mathbf{B}} \in L_f^k \text{ and } t^{\mathbf{A}} \vee t^{\mathbf{B}} \text{ exists}\}$$

*for $0 \leq k \leq n-1$.*

*Proof.* We observed already that if $r(t^{\mathbf{A}}) = r(t^{\mathbf{B}}) = k$ and $t^{\mathbf{A}} \vee t^{\mathbf{B}}$ exists, then $r(t^{\mathbf{A}} \vee t^{\mathbf{B}}) = k+1$. Thus, we have

$$L_f^{k+1}(\varphi) \supseteq \{t^{\mathbf{A}} \vee t^{\mathbf{B}} \mid t^{\mathbf{A}} \vee t^{\mathbf{B}} \in L_f^k \text{ and } t^{\mathbf{A}} \vee t^{\mathbf{B}} \text{ exists}\}$$

for $0 \leq k \leq n-1$, and we need to prove only the reverse inclusion.

Let $t^{\mathbf{C}} \in L_f^{k+1}$, where $\mathbf{C} = (c_1, \ldots, c_n)$. There exists $\ell$, $1 \leq \ell \leq n$ such that $c_\ell = *$, so $t^{\mathbf{C}}$ does not depend on $x_\ell$. If $t^{\mathbf{C}}(x_1, \ldots, x_n) = x_{i_1}^{c_{i_1}} \cdots x_{i_{n-k-1}}^{c_{i_{n-k-1}}}$, then $\ell \notin \{i_1, \ldots, i_{n-k-1}\}$ and both

$$t^{\mathbf{A}}(x_1, \ldots, x_n) = x_{i_1}^{c_{i_1}} \cdots x_\ell^0 \cdots x_{i_{n-k-1}}^{c_{i_{n-k-1}}} \text{ and}$$
$$t^{\mathbf{A}}(x_1, \ldots, x_n) = x_{i_1}^{c_{i_1}} \cdots x_\ell^1 \cdots x_{i_{n-k-1}}^{c_{i_{n-k-1}}}$$

belong to $L_f^k$. Clearly, $t^{\mathbf{C}}$ is the consensus of $t^{\mathbf{A}}$ and $t^{\mathbf{B}}$, which yields the reverse inclusion. □

Theorem 5.84 suggests that we can generate the posets of all implicants of a binary Boolean function $f : \{0,1\}^n \longrightarrow \{0,1\}$ by producing inductively the sets $L_f^0, \ldots, L_f^{n-1}$. The algorithm that implements this idea is the Quine-McCluskey algorithm, discussed next.

**Algorithm 5.85 (Quine-McCluskey Algorithm)**
**Input:** A binary Boolean function given in tabular form.
**Output:** The set $IMPL_f$ of all implicants of $f$.
**Method:** Let $L_f^0$ be the set of minterms for $f$.
For each $k$, $0 \leq k \leq n-2$, include in $L_f^{k+1}$ every term that can be obtained as a consensus of two terms from $L_f^k$.
Return the collection $\bigcup_{k=0}^{n-1} L_f^k$.

The correctness of the algorithm is an immediate consequence of Theorem 5.84.

*Example 5.86.* Consider the Boolean function $f : \{0,1\}^3 \longrightarrow \{0,1\}$ given by

| $x_1$ | $x_2$ | $x_3$ | $f(x_1, x_2, x_3)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 |

Its standard disjunctive normal form is

$$f(x_1, x_2, x_3) = t^{(0,1,1)}(x_1, x_2, x_3) \vee t^{(1,0,1)}(x_1, x_2, x_3) \\ \vee t^{(1,1,0)}(x_1, x_2, x_3) \vee t^{(1,1,1)}(x_1, x_2, x_3),$$

so the set $L_f^0$ consists of the minterms

$$t^{(0,1,1)}(x_1, x_2, x_3) = x_1^0 x_2^1 x_3^1,$$
$$t^{(1,0,1)}(x_1, x_2, x_3) = x_1^1 x_2^0 x_3^1,$$
$$t^{(1,1,0)}(x_1, x_2, x_3) = x_1^1 x_2^1 x_3^0,$$
$$t^{(1,1,1)}(x_1, x_2, x_3) = x_1^1 x_2^1 x_3^1,$$

for $(x_1, x_2, x_3) \in \{0,1\}^3$.

The Hasse diagram of $IMPL_f$ is shown in Figure 5.5. Clearly, $IMPL_f = L_f^0 \cup L_f^1$ because there is no consensus possible among any two of the implicants from $L_f^1$.

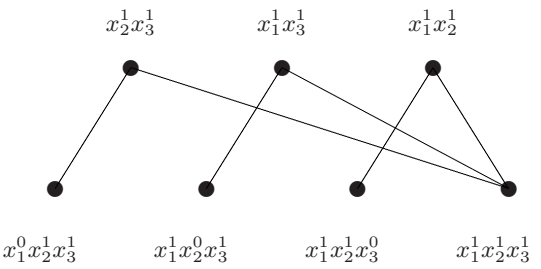



**Fig. 5.5.** Hasse diagram of $IMPL_f$.

**Definition 5.87.** *A nonempty set of terms* $T = \{t^{\boldsymbol{B}_1}, \ldots, t^{\boldsymbol{B}_m}\}$ *of implicants of a binary Boolean function* $f : \{0,1\}^n \longrightarrow \{0,1\}$ *is a* cover *of* $f$ *if* $f(x_1, \ldots, x_n) = \bigvee_{i=1}^m t^{\boldsymbol{B}_i}(x_1, \ldots, x_n)$.

$T$ *is a* minimal cover *of* $f$ *if* $T$ *is a cover of* $f$ *and no proper subset of* $T$ *is a cover of* $f$.

The set of all minterms of $f$ is clearly a cover of $f$. However, other covers may exist for $f$ that contain terms of rank that is higher than 0 and it is important to determine such simpler covers.

Since $(IMPL_f, \leq)$ is a finite poset, for every $t^{\mathbf{B}} \in IMPL_f$ there is a prime implicant $t^{\mathbf{A}}$ such that $t^{\mathbf{B}} \leq t^{\mathbf{A}}$.

**Theorem 5.88.** *Let* $f$ *be a binary Boolean function that is not the constant function* $f_0$. *A set of implicants of* $f$, $T = \{t^{\boldsymbol{B}_1}, \ldots, t^{\boldsymbol{B}_m}\}$ *is a cover of* $f$ *if and only if for every minterm* $t^{\boldsymbol{A}}$ *of* $f$ *there is an implicant* $t^{\boldsymbol{B}} \in T$ *such that* $t^{\boldsymbol{A}} \leq t^{\boldsymbol{B}_i}$.

*Proof.* Suppose that $T$ satisfies the condition of the theorem. Let $\{\mathbf{A} \in \{0,1\}^n \mid f(\mathbf{A}) = 1\} = \{\mathbf{A}_1, \ldots, \mathbf{A}_k\}$. Then, since

$$f(x_1, \ldots, x_n) = \bigvee_{1 \leq i \leq k} t^{\mathbf{A}}(x_1, \ldots, x_n) \leq \bigvee_{1 \leq l \leq m} t^{\mathbf{B}_l} \leq f(x_1, \ldots, x_n),$$

it is immediate that $T$ is a cover for $\varphi$.

Conversely, let $T$ be a cover of $f$, $T = \{t^{\mathbf{B}_1}, \ldots, t^{\mathbf{B}_m}\}$ and let $t^{\mathbf{A}}$ be a minterm, where $\mathbf{A} = (a_1, \ldots, a_n)$. Since

$$t^{\mathbf{A}}(x_1, \ldots, x_n) \leq f(x_1, \ldots, x_n) \leq \bigvee \{t^{\mathbf{B}}(x_1, \ldots, x_n) \mid t^{\mathbf{B}} \in T\},$$

it follows that there is $\mathbf{B}$ such that $t^{\mathbf{B}}(a_1, \ldots, a_n) = 1$. This implies $t^{\mathbf{A}} \leq t^{\mathbf{B}}$. □

**Corollary 5.89.** *Let $f : \{0,1\}^n \longrightarrow \{0,1\}$ be a function that is distinct from the constant function $f_0$. If $T = \{t^{\mathbf{B}_1}, \ldots, t^{\mathbf{B}_m}\}$ is a cover of $f$, and $t^{\mathbf{C}}$ is an implicant of $f$ such that $t^{\mathbf{B}_i} < t^{\mathbf{C}}$ for some $i$, $1 \leq i \leq m$, then $T' = \{t^{\mathbf{B}_1}, \ldots, t^{\mathbf{B}_{i-1}}, t^{\mathbf{C}}, t^{\mathbf{B}_{i+1}}, \ldots, t^{\mathbf{B}_m}\}$ is a cover of $f$.*

*Proof.* The statement follows immediately from Theorem 5.88. ∎

We now discuss the Quine-McCluskey systematic construction that starts with the set of prime implicants and the set of minterms of a nonzero Boolean function $f : \{0,1\}^n \longrightarrow \{0,1\}$ and yields covers of $f$ that consist of prime implicants.

Let $\mathbf{M}_f = (m_{ij})$ be a $p \times q$-matrix having one row for each prime implicant $t^{\mathbf{B}_1}, \ldots, t^{\mathbf{B}_p}$ and one column for each minterm $\{t^{\mathbf{A}_1}, \ldots, t^{\mathbf{A}_q}\}$ of $f$. Define

$$m_{ij} = \begin{cases} 1 & \text{if } t^{\mathbf{A}_j} \leq t^{\mathbf{A}_i}, \\ 0 & \text{otherwise.} \end{cases}$$

If a column of $\mathbf{M}_f$ contains a single 1, that corresponding prime implicant will be referred to as an *essential prime implicant*. Denote by $E_f$ the set of essential prime implicants for $f$. Clearly, the set $E_f$ must be contained in any cover by prime implicants of $f$.

Eliminate from $\mathbf{M}$ all essential prime implicants and the columns corresponding to the minterms they dominate.

If the set of rows of $\mathbf{M}_f$ in which a column of a minterm $t^{\mathbf{A}}$ has 1s strictly includes the set of rows in which some other column of a minterm $t^{\mathbf{A}'}$ has 1s, then eliminate column $t^{\mathbf{A}}$. Next, if among the remaining columns several have the same pattern of 1s, then retain only one of them.

Eliminate from $\mathbf{M}_f$ all rows that contain no 1s. The output consists of every minimal set of rows in $\mathbf{M}_f$ such that at least one 1 exists in these rows for every column, to each of which we add the set of essential prime implicants $E_f$.

*Example 5.90.* Let $f : \{0,1\}^4 \longrightarrow \{0,1\}$ be the binary Boolean function defined by

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $f(x_1, x_2, x_3)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 | 1 |
| 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 1 |
| 0 | 1 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 |
| 1 | 1 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 | 0 |

Starting from the minterms

$$\begin{aligned} t^{\mathbf{A}_1} &= \bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4,\ t^{\mathbf{A}_5} = \bar{x}_1x_2x_3\bar{x}_4,\\ t^{\mathbf{A}_2} &= \bar{x}_1\bar{x}_2x_3\bar{x}_4,\ t^{\mathbf{A}_6} = \bar{x}_1x_2x_3x_4,\\ t^{\mathbf{A}_3} &= \bar{x}_1x_2\bar{x}_3\bar{x}_4,\ t^{\mathbf{A}_7} = x_1\bar{x}_2x_3\bar{x}_4,\\ t^{\mathbf{A}_4} &= \bar{x}_1x_2\bar{x}_3x_4,\ t^{\mathbf{A}_8} = x_1x_2x_3\bar{x}_4, \end{aligned}$$

we have the following sets of implicants computed by using the Quine-McCluskey algorithm:

$$\begin{aligned} L_f^0 = \{&\bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4, \bar{x}_1\bar{x}_2x_3\bar{x}_4, \bar{x}_1x_2\bar{x}_3\bar{x}_4, \bar{x}_1x_2\bar{x}_3x_4,\\ &\bar{x}_1x_2x_3\bar{x}_4, \bar{x}_1x_2x_3x_4, x_1\bar{x}_2x_3\bar{x}_4, x_1x_2x_3\bar{x}_4, \},\\ L_f^1 = \{&\bar{x}_1\bar{x}_2\bar{x}_4, \bar{x}_1\bar{x}_3\bar{x}_4, \bar{x}_1x_3\bar{x}_4, \bar{x}_2x_3\bar{x}_4, \bar{x}_1x_2\bar{x}_3,\\ &\bar{x}_1x_2x_4, \bar{x}_1x_2x_3, x_2x_3\bar{x}_4, x_1x_3\bar{x}_4\},\\ L_f^2 = \{&\bar{x}_1\bar{x}_4, x_3\bar{x}_4, \bar{x}_1x_2\}. \end{aligned}$$

The prime implicants of $f$ are the terms $t^{\mathbf{B}_1} = \bar{x}_1\bar{x}_4$, $t^{\mathbf{B}_2} = x_3\bar{x}_4$, and $t^{\mathbf{B}_3} = \bar{x}_1x_2$.

The matrix $M_f$ introduced above is a $3 \times 8$ matrix:

$$M_f = \begin{pmatrix} 1&1&1&0&1&0&0&0\\ 0&1&0&0&1&0&1&1\\ 0&0&1&1&1&1&0&0 \end{pmatrix}.$$

The first, fourth, and the last three columns contain exactly one 1. Thus, all three prime implicants are essential, and they form a unique cover of prime implicants of $f$.

**Definition 5.91.** *A* partially defined Boolean function *(pdBf) on the two-element Boolean algebra is a partial function* $f : \{0,1\}^n \rightsquigarrow \{0,1\}$.

A pdBf $f : \{0,1\}^n \rightsquigarrow \{0,1\}$ is completely defined by the pair of disjoint sets

$$T_f = \{\mathbf{A} \in \mathrm{Dom}(f) \mid f(\mathbf{A}) = 1\},$$
$$F_f = \{\mathbf{A} \in \mathrm{Dom}(f) \mid f(\mathbf{A}) = 0\}.$$

**Definition 5.92.** *Let* $f : \{0,1\}^n \longrightarrow \{0,1\}$ *be a binary Boolean function and let* $i$ *be an integer such that* $1 \leq i \leq n$. *The function is* $i$-positive *if*

$$f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) \leq f(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n)$$

*for every* $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \in \{0,1\}$.

*Similarly,* $f$ *is* $i$-negative *if*

$$f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) \geq f(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n)$$

*for every* $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \in \{0,1\}$.

*The function is* $i$-monotonic *if it is either* $i$*-positive or* $i$*-negative.*

*Example 5.93.* For every $\mathbf{A} \in \{0,1,*\}^n$, the term $t^{\mathbf{A}}$ is $i$-monotonic for $1 \leq i \leq n$. Indeed, if $a_i \in \{1,*\}$, then $t^{\mathbf{A}}$ is $i$-positive; if $a_i \in \{0,*\}$, then $t^{\mathbf{A}}$ is $i$-negative.

**Theorem 5.94.** *Let* $f : \{0,1\}^n \longrightarrow \{0,1\}$ *be a binary Boolean function and let* $i$ *be an integer such that* $1 \leq i \leq n$. *If* $f$ *is* $i$*-positive, then for every prime implicant* $t^{\boldsymbol{A}}$ *of* $f$ *we have* $a_i \in \{1,*\}$, *where* $\boldsymbol{A} = (a_1, \ldots, a_n)$.

*If* $f$ *is* $i$*-negative, then* $a_i \in \{0,*\}$.

*Proof.* Suppose that $f$ is $i$-positive. Then,

$$f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) \leq f(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n)$$

for every $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \in \{0,1\}$.

Suppose that $a_i = 0$, that is,

$$t^{\mathbf{A}}(x_1, \ldots, x_n) = \cdots \wedge \bar{x}_i \wedge \cdots .$$

We claim that this implies the inequality

$$x_1^{a_1} \cdots x_{i-1}^{a_{i-1}} x_{i+1}^{a_{i+1}} \cdots x_n^{a_n} \leq f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$$

for every $x_1, \ldots, x_n \in \{0,1\}$. In other words, we have to prove that we have both

$$x_1^{a_1} \cdots x_{i-1}^{a_{i-1}} x_{i+1}^{a_{i+1}} \cdots x_n^{a_n} \leq f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)$$

and

$$x_1^{a_1} \cdots x_{i-1}^{a_{i-1}} x_{i+1}^{a_{i+1}} \cdots x_n^{a_n} \leq f(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n).$$

Since $f$ is $i$-positive, only the proof of the first inequality is necessary. The fact that $t^{\mathbf{A}}$ is an implicant of $f$ means that

$$t^{\mathbf{A}}(x_1, \ldots, x_n) \leq f(x_1, \ldots, x_n)$$

for every $x_1, \ldots, x_n \in \{0, 1\}$. Therefore,

$$\begin{aligned} t^{\mathbf{A}}(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) &= x_1^{a_1} \cdots x_{i-1}^{a_{i-1}} x_{i+1}^{a_{i+1}} \cdots x_n^{a_n} \\ &\leq f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n). \end{aligned}$$

Thus, $t^{\mathbf{B}}(x_1, \ldots, x_n) = x_1^{a_1} \cdots x_{i-1}^{a_{i-1}} x_{i+1}^{a_{i+1}} \cdots x_n^{a_n}$ is also an implicant of $f$ and, since $t^{\mathbf{A}} < t^{\mathbf{B}}$, this contradicts the fact that $t^{\mathbf{A}}$ is a prime implicant of $f$.

The second part of the theorem can be shown in a similar manner. □

## 5.6 Logical Data Analysis

Logical data analysis (LDA) is a methodology that aims to discover patterns in data using Boolean methods. LDA techniques were introduced by the Rutcor group (see [21]).

Let $\mathcal{D} = (\theta, C)$ be a pair (also referred to as a *decision system*), where $\theta = (T, H, \mathbf{r})$ is a table having the heading $H = A_1 \cdots A_n C$ and $C$ is the decision attribute. All attributes are binary, that is, we have $\mathrm{Dom}(A_1) = \cdots = \mathrm{Dom}(A_n) = \mathrm{Dom}(C) = \{0, 1\}$. The content of $\theta$ represents a sequence of observations that consists of the projections $t[A_1 \cdots A_n]$ of the tuples of $\mathbf{r}$. The component $t[C]$ of $t$ is the class of the observation $t$. The sequence of *positive observations* is

$$\mathbf{r}^+ = \{t[A_1 \cdots A_n] \text{ in } \mathbf{r} \mid t[C] = 1\};$$

the sequence of *negative observations* is

$$\mathbf{r}^- = \{t[A_1 \cdots A_n] \text{ in } \mathbf{r} \mid t[C] = 0\}.$$

It is clear that the sets $set(\mathbf{r}^+)$ and $set(\mathbf{r}^-)$ are disjoint and thus define a partial Boolean function of $n$ arguments. We refer to $\tau$ as an *observation table*.

*Example 5.95.* Consider the decision system $\mathcal{D} = ((T, A_1A_2A_3C, \mathbf{r}), C)$ given by

T

| $A_1$ | $A_2$ | $A_3$ | $C$ |
|---|---|---|---|
| 0 | 0 | 1 | 0 |
| 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 |
| 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 1 |

The sequence of positive observations is

$$\mathbf{r}^+ = ((0,0,0),(1,0,0),(1,1,0)).$$

The sequence of negative observations is

$$\mathbf{r}^- = ((0,0,1),(0,1,1),(1,1,1)).$$

Starting from a pdBf specified by two sequences of positive and negative observations, $\mathbf{r}^+ = (\mathbf{A}_1, \ldots, \mathbf{A}_p) \in \mathbf{Seq}(\{0,1\}^n)$ and $\mathbf{r}^- = (\mathbf{B}_1, \ldots, \mathbf{B}_q) \in \mathbf{Seq}(\{0,1\}^n)$, we define two corresponding binary Boolean functions $f^+$ and $f^-$ as

$$f^+(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } (x_1, \ldots, x_n) \text{ does not occur in } \mathbf{r}^-, \\ 0 & \text{otherwise,} \end{cases}$$

for $(x_1, \ldots, x_n) \in \{0,1\}^n$, and

$$f^-(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } (x_1, \ldots, x_n) \text{ does not occur in } \mathbf{r}^+, \\ 0 & \text{otherwise,} \end{cases}$$

for $(x_1, \ldots, x_n) \in \{0,1\}^n$. Clearly, we have

$$f^+(x_1, \ldots, x_n) = \bigvee \{(x_1, \ldots, x_n)^{\mathbf{A}} \mid \mathbf{A} \text{ occurs in } \mathbf{r}^-\},$$
$$f^-(x_1, \ldots, x_n) = \bigvee \{(x_1, \ldots, x_n)^{\mathbf{A}} \mid \mathbf{A} \text{ occurs in } \mathbf{r}^+\}.$$

**Definition 5.96.** *The* positive (negative) patterns of a decision system $\mathcal{D} = (\theta, C)$, *where* $\theta = (T, H, \boldsymbol{r})$, *are the prime implicants of the binary Boolean function* $f^+$ *(of the function* $f^-$*) that cover at least one minterm* $t^{\boldsymbol{A}}$, *where* $\boldsymbol{A}$ *is a positive (negative) observation of* $\tau$.

*Example 5.97.* For the positive and negative observations considered in Example 5.95, the binary Boolean functions $f^+$ and $f^-$ are given by

$$f^+(x_1, x_2, x_3, x) = \bar{x}_1\bar{x}_2\bar{x}_3 \vee \bar{x}_1 x_2 \bar{x}_3 \vee x_1\bar{x}_2\bar{x}_3 \vee x_1\bar{x}_2 x_3 \vee x_1 x_2 \bar{x}_3$$

and

$$f^-(x_1, x_2, x_3, x) = \bar{x}_1\bar{x}_2x_3 \vee \bar{x}_1x_2\bar{x}_3 \vee \bar{x}_1x_2x_3 \vee x_1\bar{x}_2x_3 \vee x_1x_2x_3$$

for $(x_1, x_2, x_3) \in \{0, 1\}^3$. The Hasse diagram of the poset of implicants $(IMPL_{f^+}, \leq)$ is shown in Figure 5.6. The prime implicants of $f^+$ are $\bar{x}_3$ and $x_1\bar{x}_2$, and they are both positive patterns. Indeed, $\bar{x}_3$ covers every positive observation, while $x_1\bar{x}_2$ covers the minterm that corresponds to the positive observation $(1, 0, 0)$.

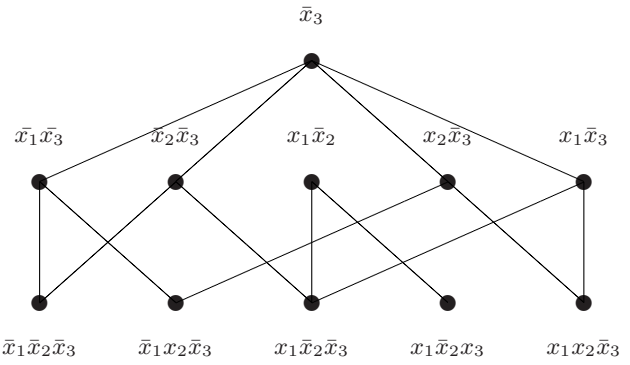



**Fig. 5.6.** Hasse diagram of $(IMPL_{f^+}, \leq)$.

A positive pattern can be regarded as a combination of values taken by a small number of variables that never appeared in a negative observation and did appear in some positive observation. Thus, if a new observation is covered by a positive pattern, this fact can be regarded as an indication that the observation is a positive one.

Next, we discuss algorithms for generating positive patterns (the negative patterns can be found using similar techniques).

Two basic approaches are described for finding positive patterns: a top-down approach and a bottom-up approach (directions are defined relative to the Hasse diagram of the poset $(\mathtt{T}(\mathcal{B}_2, n), \leq)$). In the bottom-up approach, we start with the minterms that correspond to positive observations, which are clearly positive patterns of rank 0. If one or more literals are removed from such a pattern, the resulting term may still be a pattern if it does not cover any negative examples. The process consists of a systematic removal of literals from minterms and verifying whether the resulting minterms remain positive patterns until prime patterns are reached.

The top-down approach begins with terms of rank $n - 1$; that is, with patterns that contain one literal. If such a term does not cover any negative

observations, then it is a positive pattern; otherwise, literals are added systematically until a positive pattern is obtained.

The number of positive patterns can be huge, which suggests that seeking only patterns whose rank is sufficiently high (and therefore contain few literals) is a good practical compromise.

*Example 5.98.* The Hasse diagram of the poset $(\mathtt{T}(\mathcal{B}_2, 3), \leq)$ is shown in Figure 5.7. We apply the top-down method to determine the positive patterns of the decision system introduced in Example 5.95.

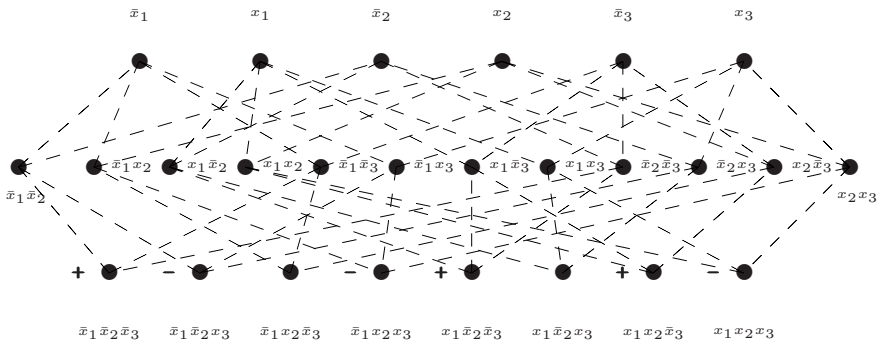



**Fig. 5.7.** Hasse diagram of the poset $(\mathtt{T}(\mathcal{B}_2, 3), \leq)$.

We begin with the minterms that correspond to positive observations,

$$t^{(0,0,0)}(x_1, x_2, x_3) = \bar{x}_1\bar{x}_2\bar{x}_3,$$
$$t^{(1,0,0)}(x_1, x_2, x_3) = x_1\bar{x}_2\bar{x}_3,$$
$$t^{(1,1,0)}(x_1, x_2, x_3) = x_1x_2\bar{x}_3.$$

The terms that can be obtained from these terms by discarding one literal are listed below.

| Original Term | Derived Term | Negative Patterns Covered |
|---|---|---|
| $\bar{x}_1\bar{x}_2\bar{x}_3$ | $\bar{x}_2\bar{x}_3$ | none |
| $\bar{x}_1\bar{x}_2\bar{x}_3$ | $\bar{x}_1\bar{x}_3$ | none |
| $\bar{x}_1\bar{x}_2\bar{x}_3$ | $\bar{x}_1\bar{x}_2$ | $\bar{x}_1\bar{x}_2x_3$ |
| $x_1\bar{x}_2\bar{x}_3$ | $\bar{x}_2\bar{x}_3$ | none |
| $x_1\bar{x}_2\bar{x}_3$ | $x_1\bar{x}_3$ | none |
| $x_1\bar{x}_2\bar{x}_3$ | $x_1\bar{x}_2$ | none |
| $x_1x_2\bar{x}_3$ | $x_2\bar{x}_3$ | none |
| $x_1x_2\bar{x}_3$ | $x_1\bar{x}_3$ | none |
| $x_1x_2\bar{x}_3$ | $x_1x_2$ | $x_1x_2x_3$ |

The preceding table yields a list of seven positive patterns of rank 1:

$$\bar{x}_2\bar{x}_3, \bar{x}_1\bar{x}_3, \bar{x}_2\bar{x}_3, x_1\bar{x}_3, x_1\bar{x}_2, x_2\bar{x}_3, x_1\bar{x}_3.$$

By discarding one literal, we have the following patterns of rank 2:

| Original Term | Derived Term | Negative Patterns Covered |
|---|---|---|
| $\bar{x}_2\bar{x}_3$ | $\bar{x}_2$ | $\bar{x}_1\bar{x}_2x_3$ |
| $\bar{x}_2\bar{x}_3$ | $\bar{x}_3$ | none |
| $\bar{x}_1\bar{x}_3$ | $\bar{x}_3$ | none |
| $\bar{x}_1\bar{x}_3$ | $\bar{x}_1$ | $\bar{x}_1\bar{x}_2x_3, \bar{x}_1x_2x_3$ |
| $\bar{x}_2\bar{x}_3$ | $\bar{x}_3$ | none |
| $\bar{x}_2\bar{x}_3$ | $\bar{x}_2$ | $\bar{x}_1\bar{x}_2x_3$ |
| $x_1\bar{x}_3$ | $\bar{x}_3$ | none |
| $x_1\bar{x}_3$ | $x_1$ | $\bar{x}_1\bar{x}_2x_3, \bar{x}_1x_2x_3$ |
| $x_1\bar{x}_2$ | $\bar{x}_2$ | $\bar{x}_1\bar{x}_2x_3$ |
| $x_1\bar{x}_2$ | $x_1$ | $x_1x_2x_3$ |
| $x_2\bar{x}_3$ | $\bar{x}_3$ | none |
| $x_2\bar{x}_3$ | $x_2$ | $\bar{x}_1x_2x_3, x_1x_2x_3$ |
| $x_1\bar{x}_3$ | $\bar{x}_3$ | none |
| $x_1\bar{x}_3$ | $x_1$ | $x_1x_2x_3$ |

The unique positive pattern of rank 2 is $\bar{x}_3$, which covers all positive patterns of rank 1 with the exception of $x_1\bar{x}_2$. Thus, we retrieve the results obtained in Example 5.97, where we used the Hasse diagram of $(IMPL_{f^+}, \leq)$.

A pattern generation algorithm is given in [21]. The algorithm gives preference to high-ranking patterns and attempts to cover every positive observation.

Data binarization is a preparatory process for LAD. Its goal is to allow the application of the Boolean methods developed in the LAD, and there are other computational benefits that follow from the binarization process.

Let $\mathcal{D} = (\theta, C)$ be a decision system where $\theta = (T, H, \mathbf{r})$ is a table having the heading $H = A_1 \cdots A_n C$. We assume that the attributes of $H$, except $C$, are nominal or numerical rather than binary. Nominal attributes have discrete domains that do not admit a natural partial order. For example, a *color* attribute having as domain the set $\{\mathsf{red}, \mathsf{white}, \mathsf{blue}\}$ is a nominal attribute. The domain of $C$ is the set $\{0, 1\}$, and we continue to refer to $\theta$ as an *observation table*.

As in Section 5.6, the content of $\theta$ represents a sequence of observations that consists of the projections $t[A_1 \cdots A_n]$ of the tuples of $\mathbf{r}$, and $t[C]$ of $t$ is the class of the observation $t$. The sequence of *positive observations* is

$$\mathbf{r}^+ = \{t[A_1 \cdots A_n] \text{ in } \mathbf{r} \mid t[C] = 1\};$$

the sequence of *negative observations* is

$$\mathbf{r}^- = \{t[A_1 \cdots A_n] \text{ in } \mathbf{r} \mid t[C] = 0\}.$$

Data binarization consists of replacing nominal or numerical attributes by binary attributes. The technique that is described here was introduced in [21].

In the case of a nominal attribute $B$ whose domain consists of the values $\{b_1, \ldots, b_p\}$, we introduce $p$ attributes $B_1, \ldots, B_p$. The $B$-component of a tuple $t[B]$ will be replaced with $p$ components corresponding to the attributes $B_1, \ldots, B_p$ such that

$$t[B_i] = \begin{cases} 1 & \text{if } t[B] = b_i, \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq i \leq p$.

For a numerical attribute $A$, we define a set of cut points. Suppose that the set of values that appear under the attribute $A$ is $\{a_1, \ldots, a_k\}$ such that $a_1 < a_2 < \cdots < a_k$. If two consecutive values $a_j$ and $a_{j+1}$ belong to two different classes, a cut point $v$ is defined as $v = \frac{a_j + a_{j+1}}{2}$. The role of cut points is to separate consecutive values of an attribute that belong to different classes.

There is no sense in choosing cut points below $\min_i a_i$ or above $\max_i a_i$ because they could not distinguish between positive or negative observations. Also, there is no reason to choose cut points between consecutive values that correspond to two positive observations or two negative observations. Therefore, we need to consider at most one cut point between any two consecutive values of $A$ that correspond to different classes.

Each cut point $v$ defines a *level binary attribute* $A_v$. The $A_v$-component of a tuple $t$ is

$$t[A_v] = \begin{cases} 0 & \text{if } t[A] < v, \\ 1 & \text{if } t[A] \geq v. \end{cases}$$

Each pair $(v, v')$ of consecutive cut points defines an *interval binary attribute* $A_{vv'}$, where the $A_{vv'}$-component of $t$ is

$$t[A_{vv'}] = \begin{cases} 0 & \text{if } v \leq t[A] < v', \\ 1 & \text{otherwise.} \end{cases}$$

*Example 5.99.* Consider the decision system $\mathcal{D} = (\theta, PlayTennis)$, where $\theta$ is the following table.

TENNIS

| Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|
| overcast | 90 | 70 | weak | 1 |
| rain | 65 | 72 | weak | 1 |
| rain | 50 | 60 | weak | 1 |
| overcast | 55 | 55 | strong | 1 |
| rain | 89 | 58 | weak | 1 |
| rain | 58 | 52 | strong | 0 |
| sunny | 75 | 75 | weak | 0 |
| rain | 77 | 77 | strong | 0 |

The attributes *Outlook* and *Wind* are nominal, while the attributes *Temperature* and *Humidity* are numerical.

Since there exist three distinct values (overcast, rain, and sunny) for the attribute *Outlook*, this attribute will be replaced with three binary attributes, $\mathrm{O}_o, \mathrm{O}_r, \mathrm{O}_s$ that correspond to these values. Similarly, the attribute *Wind* will be replaced by two binary attributes, $\mathrm{W}_w$ and $\mathrm{W}_s$.

The sequence of values for *Temperature* is shown together with the class of the observations:

$$\begin{array}{c} 50\ 55\ 58\ 65\ 75\ 77\ 89\ 90 \\ +\ +\ -\ +\ -\ -\ +\ + \end{array}$$

This requires four cut points placed at the midpoints of intervals determined by consecutive values that belong to distinct classes: $56.5, 61.5, 70, 83$.

Similarly, the sequence of values for *Humidity* is

$$\begin{array}{c} 52\ 55\ 58\ 60\ 70\ 72\ 75\ 77 \\ -\ +\ +\ +\ +\ +\ -\ - \end{array}$$

In this case, we need two cut points: 53.5 and 73.5. We use a simplified notation for binary attributes shown in the right column of the next table.

| Attribute | Simplified Notation |
|---|---|
| $\mathrm{O}_o$ | $B_1$ |
| $\mathrm{O}_r$ | $B_2$ |
| $\mathrm{O}_s$ | $B_3$ |
| $T_{56.5}$ | $B_4$ |
| $T_{61.5}$ | $B_5$ |
| $T_{70}$ | $B_6$ |
| $T_{83}$ | $B_7$ |
| $T_{56.5,61.5}$ | $B_8$ |
| $T_{61.5,70}$ | $B_9$ |
| $T_{70,83}$ | $B_{10}$ |
| $H_{53.5}$ | $B_{11}$ |
| $H_{73.5}$ | $B_{12}$ |
| $H_{53.5,73.5}$ | $B_{13}$ |
| $\mathrm{W}_w$ | $B_{14}$ |
| $\mathrm{W}_s$ | $B_{15}$ |

The binarized table is

| | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $B_5$ | $B_6$ | $B_7$ | $B_8$ | $B_9$ | $B_{10}$ | $B_{11}$ | $B_{12}$ | $B_{13}$ | $B_{14}$ | $B_{15}$ | $C$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $t_1$ | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| $t_2$ | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| $t_3$ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| $t_4$ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| $t_5$ | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| $t_6$ | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $t_7$ | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| $t_8$ | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |

The number of binary attributes can be reduced; however, the remaining attributes must allow the differentiation between positive and negative cases.

**Definition 5.100.** *Let $\mathcal{D} = (\theta, C)$ be a decision system where $\theta = (T, H, \boldsymbol{r})$ is a table having the heading $H = B_1 \cdots B_p C$ that consists of binary attributes.*

*A* support heading *for $\mathcal{D}$ is a subset $L = B_{i_1} \cdots B_{i_q}$ of $B_1 \cdots B_p$ such that if $u$ occurs in $\boldsymbol{r}^+$ and $v$ occurs in $\boldsymbol{r}^-$, then $u[L] \neq v[L]$. For any two such tuples, define their* difference set $\Delta_{\mathcal{D}}(u, v) = \{i \in H \mid u[B_i] \neq v[B_i]\}$.

*A support heading $L$ is* irredundant *if no proper subset of $L$ is a support heading of $\mathcal{D}$.*

If $L$ is a support set for a decision system $\mathcal{D} = (\theta, C)$, then $L$ must have a nonempty intersection with each set of the form $\{B_i \mid i \in \Delta_{\mathcal{D}}(u, v)\}$ for each positive example $u$ and each negative example $v$. Finding an irredundant support heading can be expressed as a discrete optimization problem, as was shown in [21].

*Example 5.101.* Let $\mathbf{y} = (y_1, \ldots, y_{14})$ be the characteristic sequence of a support heading $L$; that is,

$$y_i = \begin{cases} 1 & \text{if } B_i \in L, \\ 0 & \text{otherwise.} \end{cases}$$

The decision system introduced in Example 5.99 has five positive examples and three negative examples, so there are 15 difference sets:

$$\begin{aligned}
\Delta_{\mathcal{D}}(t_1, t_6) &= \{1, 2, 5, 6, 7, 8, 11, 13, 14, 15\}, \\
\Delta_{\mathcal{D}}(t_1, t_7) &= \{1, 3, 7, 10, 12, 13\}, \\
\Delta_{\mathcal{D}}(t_1, t_8) &= \{1, 2, 7, 10, 12, 13, 14, 15\}, \\
\Delta_{\mathcal{D}}(t_2, t_6) &= \{5, 8, 9, 11, 13, 14, 15\}, \\
\Delta_{\mathcal{D}}(t_2, t_7) &= \{2, 3, 6, 9, 10, 12, 13\}, \\
\Delta_{\mathcal{D}}(t_2, t_8) &= \{6, 9, 10, 12, 13, 14, 15\}, \\
\Delta_{\mathcal{D}}(t_3, t_6) &= \{4, 8, 11, 13, 14, 15\}, \\
\Delta_{\mathcal{D}}(t_3, t_7) &= \{2, 3, 4, 5, 6, 10, 12, 13\}, \\
\Delta_{\mathcal{D}}(t_3, t_8) &= \{4, 5, 6, 10, 12, 13, 14, 15\}, \\
\Delta_{\mathcal{D}}(t_4, t_6) &= \{1, 2, 4, 8, 11, 13\}, \\
\Delta_{\mathcal{D}}(t_4, t_7) &= \{1, 3, 4, 5, 6, 10, 12, 13, 14, 15\}, \\
\Delta_{\mathcal{D}}(t_4, t_8) &= \{1, 2, 4, 5, 6, 10, 12, 13\}, \\
\Delta_{\mathcal{D}}(t_5, t_6) &= \{5, 6, 7, 8, 11, 13, 14, 15\}, \\
\Delta_{\mathcal{D}}(t_5, t_7) &= \{2, 3, 7, 10, 12, 13\}, \\
\Delta_{\mathcal{D}}(t_5, t_8) &= \{7, 10, 12, 13, 14, 15\}.
\end{aligned}$$

The requirement that a support heading intersect each of these sets leads to 15 inequalities. For example, the requirement that $\Delta_{\mathcal{D}}(t_1, t_6) \cap L \neq \emptyset$ amounts to

$$y_1 + y_2 + y_5 + y_6 + y_7 + y_8 + y_{11} + y_{13} + y_{14} + y_{15} \geq 1.$$

Fourteen other similar inequalities can be similarly written.

Note that the set $\{13\}$ intersects all these sets, so it is a minimal support heading.

To find an irredundant support heading for a decision system $\mathcal{D} = (\theta, C)$, where $\theta = (T, H, \mathbf{r})$ is a table having the heading $H = B_1 \cdots B_p C$, we need to minimize the sum $\sum_{i=1}^{p} y_i$ subjected to restrictions of the form:

$$\sum \{y_i \mid i \in \Delta_{\mathcal{D}}(t_i, t_j)\} \geq 1$$

for every pair of tuples $(t_i, t_j)$ such that $t_i$ is a positive example and $t_j$ is a negative example.

## Exercises and Supplements

1. Consider the partially ordered sets $(P, \leq)$ and $(Q, \leq)$ whose Hasse diagrams are given in Figures 5.8(a) and (b), respectively. Determine which diagram corresponds to a lattice.

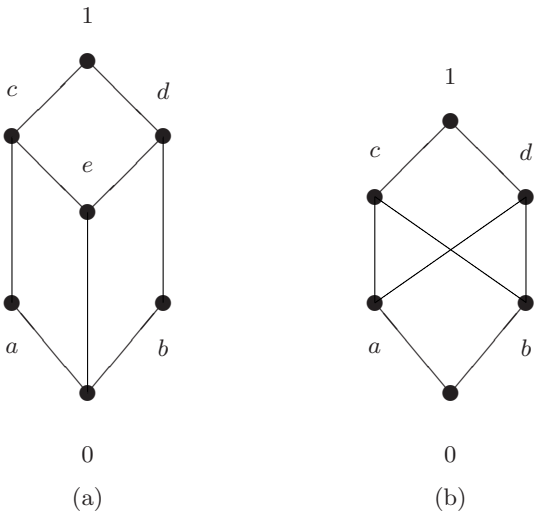



**Fig. 5.8.** Hasse diagrams of two partially ordered sets.

2. Prove that if $(L, \{\wedge, \vee\})$ is a lattice, then for every finite, nonempty subset $K$ of $L$, $\inf K$ and $\sup K$ exist.
3. Prove that every chain is a lattice.
4. Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a lattice and let $x$ and $y$ be two elements of $L$. Prove that the least sublattice of $\mathcal{L}$ that contains $x$ and $y$ is $\{x, y, x \wedge y, x \vee y\}$.

Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a lattice. A nonempty subset $I$ of $L$ is an *ideal* of $\mathcal{L}$ if $x \vee y \in I$ holds if and only if both $x \in I$ and $y \in I$. A *filter* of $\mathcal{L}$ is a nonempty subset $F$ of $L$ such that $x \wedge y \in F$ if and only if $x \in F$ and $y \in F$.

5. Prove that a set $K$ is an ideal of the lattice $\mathcal{L} = (L, \{\wedge, \vee\})$ if and only if $x \in K$ and $y \in K$ imply $x \vee y \in K$, and $x \in K$ and $t \leq x$ imply $t \in K$.
6. Prove that a set $K$ is a filter of the lattice $\mathcal{L} = (L, \{\wedge, \vee\})$ if and only if $x \in K$ and $y \in K$ imply $x \wedge y \in I$, and $x \in K$ and $t \geq x$ imply $t \in K$.
7. Prove that, for every element $x$ of a lattice $\mathcal{L} = (L, \{\wedge, \vee\})$, the set $I_x = \{t \in L \mid t \leq x\}$ is an ideal and the set $F_x = \{t \in L \mid x \leq t\}$ is a filter. They are referred to as the *principal ideal* of $x$ and the *principal filter* of $x$.
8. Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a lattice and let $\mathbf{A} = (a_{ij})$ be an $m \times n$ matrix of elements of $L$.
   a) Prove the following generalization of the minimax inequality (see Exercise 22 of Chapter 2):
   $$\bigvee_j \bigwedge_i a_{ij} \leq \bigwedge_i \bigvee_j a_{ij}$$
   b) Suppose that $L$ has the least element 0. Write the inequality that follows from the application of the minimax inequality to the matrix
   $$\mathbf{A} = \begin{pmatrix} 0 & a & b \\ b & 0 & c \\ a & c & 0 \end{pmatrix}.$$
9. Prove the following generalization of Theorem 5.40. In a distributive lattice $\mathcal{L} = (L, \{\wedge, \vee\})$, the equalities $x \vee y = x \vee z$ and $x \wedge y = x \vee z$ imply $y = z$. Conversely, if $x \vee y = x \vee z$ and $x \wedge y = x \vee z$ imply $y = z$ for all $x, y, z \in L$, then $L$ is distributive.
10. Prove that every lattice having four elements is distributive.
11. Prove that a lattice $\mathcal{L} = (L, \{\wedge, \vee\})$ is modular if and only if
    $$x \wedge (y \vee (x \wedge z)) = (x \wedge y) \vee (x \wedge z)$$
    for every $x, y, z \in L$.
12. Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a lattice that has the least element 0. If the set $\{t \in L \mid x \vee t = 0\}$ has a largest element $x^*$, then we say that $x^*$ is the *pseudocomplement* of $x$. If every element of $L$ has a pseudocomplement, then we say that $L$ is a *pseudocomplemented lattice*.

a) Prove that any pseudocomplemented lattice has a largest element.
b) Prove that if $\mathfrak{L}$ is a chain having the least element 0 and the largest element 1, then $L$ is pseudocomplemented.
c) Prove that $x \leq x^{**}$ and $x^* = x^{***}$ for $x \in L$.
d) Prove that $(x \wedge y)^{**} = x^{**} \wedge y^{**}$ for $x, y \in L$.

13. Let $(S, \leq)$ be a poset, $x$ be an element of $S$, and $I_x = \{s \in S \mid s \leq x\}$.
a) Prove that for every $x \in S$ the set $I_x = \{s \in S \mid s \leq x\}$ is an order ideal of $(S, \leq)$. This is *the principal order ideal of* $x$.
b) Let $\mathfrak{I}_p(S, \leq)$ be the collection of principal order ideals of $(S, \leq)$ and let $f : S \longrightarrow \mathfrak{I}_p(S, \leq)$ be the mapping defined by $f(x) = I_x$ for $x \in S$. Prove that $f$ is a monotonic injection.
c) Let $T$ be a subset of $S$. Prove that if $\sup T$ ($\inf T$) exists in $(S, \leq)$, then $\sup\{I_x \mid x \in T\}$ in $(\mathfrak{I}_p(S, \leq), \subseteq)$ is $I_{\sup T}$ ($\inf\{I_x \mid x \in T\}$ in $(\mathfrak{I}_p(S, \leq), \subseteq)$ is $I_{\inf T}$).
d) Prove that the poset of principal order ideals of $S$, $(\mathfrak{I}_p(S, \leq), \subseteq)$ is a complete lattice.

14. Let $(L_1, \leq)$ and $(L_2, \leq)$ be two complete lattices and let $f : L_1 \longrightarrow L_2$ be a monotonic function between these posets. Prove that

$$f\left(\bigvee K\right) \geq \bigvee f(K),$$
$$f\left(\bigwedge K\right) \leq \bigwedge f(K),$$

for every subset $K$ of $L_1$.

15. Let $(L, \leq)$ be a complete lattice and let $f : L \longrightarrow L$ be a monotonic mapping. Prove that there exists $x \in L$ such that $f(x) = x$ (Tarski's fixed-point theorem).

**Solution:** Let $T = \{x \in L \mid x \leq f(x)\}$ and $t = \sup T$. Since $x \leq t$, we have $f(x) \leq f(t)$ for every $x \in T$, so $x \leq f(x) \leq f(t)$. This implies $t \leq f(t)$, so $t \in T$. Therefore, $f(t) \leq f(f(t))$, so $f(t) \in T$, which implies $f(t) \leq t$. This shows that $t = f(t)$.

16. Let $S$ and $T$ be two sets and let $f : S \longrightarrow T$, $g : T \longrightarrow S$ be two injective functions. Define the function $F : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ as $F(U) = S - g(T - f(U))$ for every $U \in \mathcal{P}(S)$.
a) Prove that $F$ is a monotonic mapping between the complete lattices $(\mathcal{P}(S), \subseteq)$ and $(\mathcal{P}(T), \subseteq)$.
b) Let $U_0 \subseteq S$ be a fixed point of $F$. Define the function $h : S \longrightarrow T$ by

$$h(x) = \begin{cases} f(x) & \text{if } x \in U_0, \\ y & \text{if } x \notin U_0 \text{ and } g(y) = x. \end{cases}$$

Show that $h$ is well-defined and, moreover, is a bijection. (The existence of a bijection $h$ between $S$ and $T$ when the injections $f$ and $g$ exist is known as the *Schröder-Bernstein theorem*.)

17. Prove that $f : B^n \longrightarrow B$ is a Boolean function if and only if

$$(x_1, \ldots, x_n)^{\mathbf{A}} f(x_1, \ldots, x_n) = (x_1, \ldots, x_n)^{\mathbf{A}} f(\mathbf{A})$$

for every $\mathbf{A} \in \{0, 1\}^n$.

18. Prove that there are $2^{n-k}\binom{n}{k}$ $n$-ary terms of rank $k$.

19. A *Horn term* is a term $t^{\mathbf{A}} : B^n \longrightarrow B$ such that $\mathbf{A}$ contains at most one 0. Prove that if the consensus $t^{\mathbf{C}}$ of the Horn $t^{\mathbf{A}}$, $t^{\mathbf{B}}$ exists, then $t^{\mathbf{C}}$ is a Horn term.

20. Let $\mathcal{B} = (B, \{\wedge, \vee, ^-, 0, 1\})$ be a Boolean algebra and let $f : B^n \longrightarrow B$ be a Boolean function. Prove that, for every $b \in B$, there exists a Boolean function $f_{(b)} : B^n \longrightarrow B$ such that

$$b \wedge f(x_0, \ldots, x_{n-1}) = f_{(b)}(b \wedge x_0, \ldots, b \wedge x_{n-1})$$

for $(x_0, \ldots, x_{n-1}) \in B^n$.

**Hint:** The argument is by induction on the definition of Boolean functions.

21. Let $S$ be a set and let $\pi$ be a partition of $S$. Prove that the collection of $\pi$-saturated subsets of $S$ is a Boolean subalgebra $S_\pi$ of the Boolean algebra $(\mathcal{P}(S), \{\cap, \cup, ^-, \emptyset, S\})$.

22. Let $\mathcal{C}$ be a finite collection of subsets of a set $S$. Prove that the subalgebra of $(\mathcal{P}(S), \{\cap, \cup, ^-, \emptyset, S\})$ generated by $\mathcal{C}$ coincides with the collection of all $\pi_{\mathcal{C}}$-saturated sets, where $\pi_{\mathcal{C}}$ is the partition defined in Supplement 6 of Chapter 1. Further, show that the atoms of this subalgebra are the blocks of the partition $\pi_{\mathcal{C}}$.

23. Let $S$ and $T$ be two sets and let $f : S \longrightarrow T$ be a mapping.
    a) Prove that the function $F : \mathcal{P}(T) \longrightarrow \mathcal{P}(S)$ defined by $F(V) = f^{-1}(V)$ for $V \in \mathcal{P}(T)$ is a Boolean algebra morphism between the algebras $(\mathcal{P}(T), \{\cap, \cup, ^-, \emptyset, T\})$ and $(\mathcal{P}(S), \{\cap, \cup, ^-, \emptyset, S\})$.
    b) Let $\mathcal{D} = \{D_1, \ldots, D_r\}$ be a finite collection of subsets of $T$ and let $\mathcal{C} = \{F(D) \mid D \in \mathcal{D}\}$ be the corresponding finite collection of subsets of $C$.
    If $\pi_{\mathcal{D}}$ is the partition of $T$ associated to $\mathcal{D}$, then prove that for any block $B$ of this partition, $F(B)$ is either the empty set or a block of the partition $\pi_{\mathcal{C}}$ of $S$, and each block of the partition $\pi_{\mathcal{C}}$ is of the form $F(B)$. Further, if $f$ is a surjective mapping, then $F(B)$ is always a block of $\pi_{\mathcal{C}}$.

**Solution:** The first part is a consequence of Theorems 1.65 and 1.67. For the second part, let

$$D_{a_1,\ldots,a_r} = D_1^{a_1} \cap \cdots \cap D_r^{a_r}$$

be an atom of $\pi_{\mathcal{D}}$ for some $(a_1, \ldots, a_r)$. Note that $F(D_{a_1,\ldots,a_r}) = C_1^{a_1} \cap \cdots \cap C_r^{a_r}$, where $C_i = f^{-1}(D_i) = F(D_i)$ for $1 \leq i \leq r$. If this intersection is nonempty, then it is clearly a block of $\pi_{\mathcal{D}}$.

If $f$ is surjective, the preimage of any nonempty set $f^{-1}(D_{a_1,\ldots,a_r})$ is nonempty and therefore a block of $\pi_{\mathcal{C}}$.

24. Let $F : \{0,1\}^n \longrightarrow \{0,1\}^n$ be a function. Prove that there exists a bijection $G : \{0,1\}^n \longrightarrow \{0,1\}^n$ such that $G(\mathbf{x}) \wedge \mathbf{x} = F(\mathbf{x}) \wedge \mathbf{x}$ for every $\mathbf{x} \in \{0,1\}^n$.

**Solution:** Without loss of generality, we may assume that $F(\mathbf{x}) \leq \mathbf{x}$ for $\mathbf{x} \in \{0,1\}^n$. Thus, we need to prove the existence of $G$ such that $G(\mathbf{x}) \wedge \mathbf{x} = F(\mathbf{x})$.

For $\mathbf{x} \subseteq \{0,1\}^n$ let $K_F(\mathbf{x}) = \{\mathbf{u} \in \{0,1\}^n \mid \mathbf{u} \wedge \mathbf{x} = F(\mathbf{x})\}$. If $X \subseteq \{0,1\}^n$, define $K_F(X) = \bigcup\{K(\mathbf{x}) \mid \mathbf{x} \in X\}$. Then, we should have $G(\mathbf{x}) \in K_F(\mathbf{x})$ for each $\mathbf{x} \in \{0,1\}^n$. To obtain the result it suffices to show that, for every $X$, we have $|X| \leq |K_F(X)|$ because this would imply that there is a bijection $G$ such that $G(\mathbf{x}) \in K_F(\mathbf{x})$.

Note that if $F(\mathbf{x}) = \mathbf{x}$, then $K_F(\mathbf{x}) = \{\mathbf{u} \in \{0,1\}^n \mid \mathbf{u} \geq \mathbf{x}\}$, so $\mathbf{x} \in K_F(\mathbf{x})$ for every $\mathbf{x} \in \{0,1\}^n$, which implies $|X| \leq |K_F(X)|$.

For $\mathbf{x} = (x_1, \ldots, x_n)$ and $F(\mathbf{x}) = (y_1, y_2, \ldots, y_n)$, define $F_1 : \{0,1\}^n \longrightarrow \{0,1\}^n$ by modifying the first component of $F(\mathbf{x})$ as

$$F_1(\mathbf{x}) = (x_1, y_2, \ldots, y_n).$$

If $\mathbf{u} = (u_1, u_2, \ldots, u_n) \in \{0,1\}^n$, denote by $\mathbf{u}_{[0]}$ and $\mathbf{u}_{[1]}$ the $n$-tuples $\mathbf{u}_{[0]} = (0, u_2, \ldots, u_n)$ and $\mathbf{u}_{[1]} = (1, u_2, \ldots, u_n)$.

We claim that $|K_{F_1}(X)| \leq |K_F(X)|$. To prove this inequality, it suffices to show that $|K_{F_1}(X) \cap \{\mathbf{u}_{[0]}, \mathbf{u}_{[1]}\}| \leq |K_F(X) \cap \{\mathbf{u}_{[0]}, \mathbf{u}_{[1]}\}|$ for every $\mathbf{u} \in \{0,1\}^n$.

If $\mathbf{u}_{[0]} \in K_{F_1}(X)$, then $\{\mathbf{u}_{[0]}, \mathbf{u}_{[1]}\} \subseteq K_F(X)$. Indeed, $\mathbf{u}_{[0]} \in K_{F_1}(X)$ implies $\mathbf{u}_{[0]} \wedge \mathbf{x} = F_1(\mathbf{x})$ for some $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in X$, which yields $x_1 = 0$. Since $F(\mathbf{x}) \leq \mathbf{x}$, it follows that $(F(\mathbf{x}))_1 = 0$ and $(F(\mathbf{x}))_1 = (F_1(\mathbf{x}))_1$. Thus, $\mathbf{u}_{[1]} \wedge \mathbf{x} = \mathbf{u}_{[0]} \wedge \mathbf{x} = F(\mathbf{x})$, so $\{\mathbf{u}_{[0]}, \mathbf{u}_{[1]} \subseteq K_F(X)\}$. If $\mathbf{u}_{[1]} \in\in F_1(X)$ and $\mathbf{u}_{[0]} \notin F_1(X)$, then $\{\mathbf{u}_{[0]}, \mathbf{u}_{[1]}\} \cap K_F(X) \neq \emptyset$. Under these assumptions, there exists $\mathbf{x} \in X$ such that $\mathbf{u}_{[1]} \wedge \mathbf{x} = F_1(\mathbf{x})$ and $\mathbf{u}_{[0]} \wedge \mathbf{x} \neq F_1(\mathbf{x})$. Note that $(\mathbf{u}_{[0]} \wedge \mathbf{x})_i = (\mathbf{u}_{[1]} \wedge \mathbf{x})_i = (F_1(\mathbf{x}))_i = (F(\mathbf{x}))_i$ for $2 \leq i \leq n$. Also, we have $(F(\mathbf{x}))_1 = 0 = (\mathbf{u}_{[0]} \wedge \mathbf{x})_1$ or $(F(\mathbf{x}))_1 = 1 = (\mathbf{u}_{[1]} \wedge \mathbf{x})_1$. Thus, either $\mathbf{u}_{[0]} \wedge \mathbf{x} = F(\mathbf{x})$ or $\mathbf{u}_{[1]} \wedge \mathbf{x} = F(\mathbf{x})$.

The treatment applied to the first coordinate can now be repeated for the second component starting from $F_1$ to produce a function $F_2 : \{0,1\}^n \longrightarrow \{0,1\}^n$ such that $|K_{F_2}(X)| \leq |K_{F_1}(X)|$, etc. After $n$ steps, we reach a function $F_n$ such that $F_n(\mathbf{x}) = \mathbf{x}$. We have $K_{F_n}(\mathbf{x}) = \{\mathbf{u} \in \{0,1\}^n \mid \mathbf{u} \geq \mathbf{x}\}$, so $X \subseteq K_{F_n}(X)$, which implies $|X| \leq |K_{F_n}(X)| \leq |K_F(X)|$.

25. Let $S$ be a finite set and let $U$ be a subset of $S$. Prove that there exists a bijection $G_U : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ such that $G_U(T) \wedge T = T \cap U$ for every subset $T$ of $S$.

## Bibliographical Comments

The books [114, 115] are essential references for Boolean algebra and Boolean functions. A comprehensive reference on pseudo-Boolean functions is the monograph [63]. Basic references for logical data analysis are [21] and [20]. Supplement 24 appears in a slightly different form in [108].

# 6

# Topologies and Measures

## 6.1 Introduction

Topology is an area of mathematics that investigates both the local and the global structure of space. The term "topology" is derived from the Greek words $\tau\acute{o}\pi o\zeta$ (*topos*, place) and $\lambda\acute{o}\gamma o\zeta$ (*logos*, reason) and was introduced in [90]. We present in this chapter an introduction to point-set topology that is important for a subsequent discussion of various notions of dimensions of sets. Data mining makes use of topology in formulating searching algorithms that take into account the local properties of data sets.

## 6.2 Topologies

The term "topology" is used both to designate a mathematical discipline and to name the fundamental notion of this discipline, which is introduced next.

**Definition 6.1.** *A* topology *on a set $S$ is a family $\mathcal{O}$ of subsets of $S$ that satisfies the following conditions.*
(i) $\emptyset \in \mathcal{O}$ *and* $S \in \mathcal{O}$.
(ii) *For every collection $\mathcal{C}$ such that $\mathcal{C} \subseteq \mathcal{O}$, $\bigcup \mathcal{C} \in \mathcal{O}$.*
(iii) *If $\mathcal{D}$ is a finite collection and $\mathcal{D} \subseteq \mathcal{O}$, then $\bigcap \mathcal{D} \in \mathcal{O}$.*
*The sets that belong to $\mathcal{O}$ are referred to as the* open sets *of the topology $\mathcal{O}$. The pair $(S, \mathcal{O})$ will be referred to as a* topological space.

It is easy to see that Part (iii) of Definition 6.1 is equivalent to
(iii′) if $U, U' \in \mathcal{O}$, then $U \cap U' \in \mathcal{O}$.

Actually, the first condition of Definition 6.1 is superfluous. Indeed, since we deal here with collections of subsets of $S$, by Part (iii) of the definition, the intersection of the empty collection of subsets of $S$ belongs to $\mathcal{O}$, and this intersection is $S$. On the other hand, by Part (ii), the union of the empty collection (which is the empty set) belongs to $\mathcal{O}$, so Part (i) is a consequence of the remaining parts of the definition.

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_6, © Springer-Verlag London Limited 2008

*Example 6.2.* The pair $(S, \mathcal{P}(S))$ is a topological space. The topology $\mathcal{P}(S)$ is known as the *discrete topology*.

The collection $\{\emptyset, S\}$ is the *indiscrete topology*.

*Example 6.3.* The pair $(\emptyset, \{\emptyset\})$ is a topological space as the reader can easily verify. We refer to $(\emptyset, \{\emptyset\})$ as the *empty topological space*.

*Example 6.4.* Let $\mathcal{O}$ be the collection of subsets of $\mathbb{R}$ defined by $L \in \mathcal{O}$ if for every $x \in L$ there exists $\epsilon \in \mathbb{R}_{>0}$ such that $|u - x| < \epsilon$ implies $u \in L$. We claim that $\mathcal{O}$ is a topology on $\mathbb{R}$.

Indeed, it is immediate that $\emptyset$ and $\mathbb{R}$ belong to $\mathcal{O}$.

Let $\mathcal{C}$ be such that $\mathcal{C} \subseteq \mathcal{O}$ and let $x \in \bigcup \mathcal{C}$. There exists $L \in \mathcal{C}$ such that $x \in L$ and, therefore, by the definition of $\mathcal{O}$, there is $\epsilon > 0$ such that $|u-x| < \epsilon$ implies $u \in L$. Thus, $u \in \bigcup \mathcal{C}$, so $\bigcup \mathcal{C} \in \mathcal{O}$.

Suppose now that $\mathcal{D}$ is a finite subcollection of $\mathcal{O}$, $\mathcal{D} = \{D_1, \ldots, D_n\}$, and let $x \in \bigcap \mathcal{D}$. Since $x \in D_i$ for $1 \leq i \leq n$, there exists $\epsilon_1, \ldots, \epsilon_n$ such that $|u - x| < \epsilon_i$ implies $u \in D_i$ for every $i$, $1 \leq i \leq n$. Therefore, by defining $\epsilon = \min\{\epsilon_i \mid 1 \leq i \leq n\}$, it follows that $|x - u| \leq \epsilon$ implies $u \in \bigcap \mathcal{D}$, which proves that $\bigcap \mathcal{D} \in \mathcal{O}$. We conclude that $\mathcal{O}$ is a topology on $\mathbb{R}$. This topology is called the *usual topology* on $\mathbb{R}$. Unless stated otherwise, we assume that the set of real numbers is equipped with the usual topology.

*Example 6.5.* Example 6.4 can be extended to $\mathbb{R}^n$ by defining the collection of sets $\mathcal{O}$ as consisting of subsets $L$ of $\mathbb{R}^n$ such that for every $\mathbf{x} \in L$ there exists $\epsilon \in \mathbb{R}_{>0}$ such that $d_2(\mathbf{u}, \mathbf{x}) < \epsilon$ implies $\mathbf{u} \in L$. It is easy to verify that $(\mathbb{R}^n, \mathcal{O})$ is a topological space.

For each topology $\mathcal{O}$ on a set $S$, we define the collection of *closed sets* as

$$closed(\mathcal{O}) = \{S - X \mid X \in \mathcal{O}\}$$

and the *collection of neighborhoods* of an element $x$ of $S$ as

$$neigh_x(\mathcal{O}) = \{U \in \mathcal{P}(S) | \text{ there is } W \in \mathcal{O} \text{ such that } x \in W \subseteq U\}.$$

**Theorem 6.6.** *Let $(S, \mathcal{O})$ be a topological space. The following statements hold:*

(i) *$\emptyset$ and $S$ are closed sets.*

(ii) *For every collection $\mathcal{C}$ of closed sets, $\bigcap \mathcal{C}$ is a closed set.*

(iii) *For every finite collection $\mathcal{D}$ of closed sets, $\bigcup \mathcal{D}$ is a closed set.*

*Proof.* This is an immediate consequence of Definition 6.1. □

## 6.3 Closure and Interior Operators in Topological Spaces

Theorem 6.6 implies that for every topological space $(S, \mathcal{O})$ the collection $closed(\mathcal{O})$ of closed sets is a closure system on $S$. For the closure operator

attached to this closure system denoted by $\mathbf{K}_{S,\mathcal{O}}$, we have the supplementary property:

$$\mathbf{K}_{S,\mathcal{O}}(H \cup L) = \mathbf{K}_{S,\mathcal{O}}(H) \cup \mathbf{K}_{S,\mathcal{O}}(L) \tag{6.1}$$

for all subsets $H, L$ of $S$.

Since $H, L \subseteq H \cup L$, we have $\mathbf{K}_{S,\mathcal{O}}(H) \subseteq \mathbf{K}_{S,\mathcal{O}}(H \cup L)$ and $\mathbf{K}_{S,\mathcal{O}}(L) \subseteq \mathbf{K}_{S,\mathcal{O}}(H \cup L)$ due to the monotonicity of $\mathbf{K}_{S,\mathcal{O}}$. Therefore,

$$\mathbf{K}_{S,\mathcal{O}}(H) \cup \mathbf{K}_{S,\mathcal{O}}(L) \subseteq \mathbf{K}_{S,\mathcal{O}}(H \cup L).$$

To prove the reverse inclusion, note that the set $\mathbf{K}_{S,\mathcal{O}}(H) \cup \mathbf{K}_{S,\mathcal{O}}(L)$ is a closed set by the third part of Theorem 6.6 and $H \cup L \subseteq \mathbf{K}_{S,\mathcal{O}}(H) \cup \mathbf{K}_{S,\mathcal{O}}(L)$. Therefore, the closure of $H \cup L$ is a subset of $\mathbf{K}_{S,\mathcal{O}}(H) \cup \mathbf{K}_{S,\mathcal{O}}(L)$, so $\mathbf{K}_{S,\mathcal{O}}(H \cup L) \subseteq \mathbf{K}_{S,\mathcal{O}}(H) \cup \mathbf{K}_{S,\mathcal{O}}(L)$, which implies Equality (6.1).

Also, note that $\mathbf{K}_{S,\mathcal{O}}(\emptyset) = \emptyset$ because the empty set itself is closed.

If there is no risk of confusion, we will denote the closure operator $\mathbf{K}_{S,\mathcal{O}}$ simply by $\mathbf{K}$.

Note that Equality (6.1) is satisfied for every $H, L \in \mathcal{P}(S)$ if and only if the union of two $\mathbf{K}$-closed sets is $\mathbf{K}$-closed. Indeed, suppose that Equality (6.1) is satisfied, and let $U$ and $V$ be two $\mathbf{K}$-closed sets. Since $U = \mathbf{K}(U)$ and $V = \mathbf{K}(V)$, it follows that $U \cup V = \mathbf{K}(U) \cup \mathbf{K}(V) = \mathbf{K}(U \cup V)$, which shows that $U \cup V$ is $\mathbf{K}$-closed. Conversely, suppose that the union of two $\mathbf{K}$-closed sets is $\mathbf{K}$-closed. Then, $\mathbf{K}(U) \cup \mathbf{K}(V)$ is $\mathbf{K}$-closed and contains $U \cup V$. Therefore, $\mathbf{K}(U \cup V) \subseteq \mathbf{K}(U) \cup \mathbf{K}(V)$. The reverse equality follows from the monotonicity of $\mathbf{K}$.

**Theorem 6.7.** *Let $S$ be a set and let $\boldsymbol{K} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ be a closure operator that satisfies Equality (6.1) for every $H, L \in \mathcal{P}(S)$ and $\boldsymbol{K}(\emptyset) = \emptyset$. The collection $\mathcal{O}_{\boldsymbol{K}} = \{S - U \mid U \in \mathcal{C}_{\boldsymbol{K}}\}$ is a topology on $S$.*

*Proof.* We have $\mathbf{K}(S) = S$, so both $\emptyset$ and $S$ are $\mathbf{K}$-closed sets, which implies $\emptyset, S \in \mathcal{O}_{\mathbf{K}}$.

Suppose that $\mathcal{C} = \{L_i \mid i \in I\} \subseteq \mathcal{O}_{\mathbf{K}}$. Since $S - L_i \in \mathcal{C}_{\mathbf{K}}$, it follows that $\bigcap\{S - L_i \mid i \in I\} = S - \bigcup_{i \in I} L_i \in \mathcal{C}_{\mathbf{K}}$. Thus, $\bigcup_{i \in I} L_i \in \mathcal{O}_{\mathbf{K}}$.

Finally, suppose that $\mathcal{D} = \{D_1, \ldots, D_n\}$ is a finite collection of subsets such that $\mathcal{D} \subseteq \mathcal{O}_{\mathbf{K}}$. Since $S - D_i \in \mathcal{C}_{\mathbf{K}}$ we have $S - \bigcup_{i=1}^{n} D_i = \bigcap_{i=1}^{n}(S - D_i) \in \mathcal{C}_{\mathbf{K}}$, hence $\bigcup_{i=1}^{n} D_i \in \mathcal{O}_{\mathbf{K}}$. This proves that $\mathcal{O}_{\mathbf{K}}$ is indeed a topology. □

**Theorem 6.8.** *Let $(S, \mathcal{O})$ be a topological space and let $U$ and $W$ be two subsets of $S$. If $U$ is open and $U \cap W = \emptyset$, then $U \cap \boldsymbol{K}(W) = \emptyset$.*

*Proof.* $U \cap W = \emptyset$ implies $W \subseteq S - U$. Since $U$ is open, the set $S - U$ is closed, so $\mathbf{K}(W) \subseteq \mathbf{K}(S - U) = S - U$. Therefore, $U \cap \mathbf{K}(W) = \emptyset$. □

Often, we shall use the contrapositive of this statement: if $U$ is an open set such that $U \cap \mathbf{K}(W) \neq \emptyset$ for some set $W$, then $U \cap W \neq \emptyset$.

*Example 6.9.* In the topological space $(\mathbb{R}, \mathcal{O})$, every open interval $(a, b)$ with $a < b$ is an open set. Indeed, if $x \in (a, b)$ and $|x-u| < \epsilon$, where $\epsilon = \frac{1}{2}\min\{|x-a|, |x-b|\}$, then $u \in (a, b)$. A similar argument shows that the half-lines $(b, +\infty)$ and $(-\infty, a)$ are open sets for $a, b \in \mathbb{R}$. Therefore, $(-\infty, a) \cup (b, +\infty)$ is an open set which implies that its complement, the interval $[a, b]$, is closed. Also, $(-\infty, b]$ and $[a, \infty)$ are closed sets (as complements of the open sets $(b, \infty)$ and $(a, \infty)$, respectively).

Open sets of the topological space $(\mathbb{R}, \mathcal{O})$, where $\mathcal{O}$ is the usual topology on the set of real numbers have the following useful characterization.

**Theorem 6.10.** *A subset $U$ of $\mathbb{R}$ is open in the topological space $(\mathbb{R}, \mathcal{O})$ if and only if it may be written as a union of a countable collection of disjoint open intervals.*

*Proof.* Since every open interval (finite or not) is an open set, it follows that the union of a countable collection of disjoint open intervals is open.

To prove the converse, let $U$ be an open set. Note that $U$ can be written as a union of open intervals since for each $x \in U$ there exists $\epsilon > 0$ such that $x \in (x - \epsilon, x + \epsilon) \subseteq U$.

Define the relation $\theta_U$ on the set $U$ by $x\theta_U y$ if there exist $a, b \in \mathbb{R}$ such that $\{x, y\} \subseteq (a, b) \subseteq U$, where $(a, b)$ is the open interval determined by $a, b$. We claim that $\theta_U$ is an equivalence relation on $U$.

Since $U$ is open, $x \in U$ implies the existence of a positive number $\epsilon$ such that $\{x\} \subseteq (x - \epsilon, x + \epsilon) \subseteq U$ for every $x \in U$, so $\theta_U$ is reflexive. The symmetry of $\theta_U$ is immediate. To prove its transitivity, let $x, y, z \in U$ be such that $x\theta_U z$ and $z\theta_U y$. There are $a, b, c, d \in \mathbb{R}$ such that $\{x, z\} \subseteq (a, b) \subseteq U$ and $\{z, y\} \subseteq (c, d) \subseteq U$. Since $z \in (a, b) \cap (c, d)$, it follows that $(a, b) \cup (c, d)$ is an interval $(e, e')$ such that $\{x, y\} \subseteq (e, e') \subseteq U$, which shows that $x\theta_U y$. Thus, $\theta_U$ is an equivalence on $U$.

We claim that each equivalence class $[x]_{\theta_U}$ is an open interval or a set of the form $(a, +\infty)$ or a set of the form $(-\infty, b)$. Indeed, suppose that $u, v \in [x]_{\theta_U}$ (that is, $u\theta_U x$ and $v\theta_U x$) and that $t \in (u, v)$. We now prove that $t\theta_U x$.

There are two open intervals $(a, b)$ and $(c, d)$ such that $\{u, x\} \subseteq (a, b) \subseteq U$ and $\{x, v\} \subseteq (c, d) \subseteq U$. Again, $(a, b) \cup (c, d)$ is an open interval $(e, e')$ and we have $(u, v) \subseteq (e, e') \subseteq U$. Thus, if $[x]_{\theta_U}$ contains two numbers $u$ and $v$, it also contains the interval $(u, v)$ determined by these numbers.

To prove that $[x]_{\theta_U}$ has the desired form, we will prove that this set has no least element and no greatest element. Suppose that $[x]_{\theta_U}$ has a least element $y$. Then, there exist $a$ and $b$ such that $a < y < x < b$ and $(a, b) \subseteq U$. Since $y$ is supposed to be the least element of $[x]_{\theta_U}$, if $a < z < y$, we have $z \notin [x]_{\theta_U}$. This contradicts $y\theta_U z$ and $y\theta_U x$. In a similar manner, it is possible to show that $[x]_{\theta_U}$ has no largest element.

Finally, we prove that the partition that corresponds to $\theta_U$ is countable. Select a rational number $r_x \in [x]_{\theta_U} \cap \mathbb{Q}$. Since the equivalence classes $[x]_{\theta_U}$ are pairwise disjoint, it follows that $[x]_{\theta_U} \neq [y]_{\theta_U}$ implies $r_x \neq r_y$. Thus, we

have an injection $r : U/\theta_U \longrightarrow \mathbb{Q}$ given by $r([x]_{\theta_U}) = r_x$ for $x \in U$. By Theorem 1.131, the set $U/\theta_U$ is countable. □

*Example 6.11.* Let $S$ be an infinite set. The family of sets

$$\mathcal{O} = \{\emptyset\} \cup \{L \in \mathcal{P}(S) \mid S - U \text{ is finite}\}$$

is a topology on $S$. We shall refer to $\mathcal{O}$ as the *cofinite topology* on $S$.

Note that both $\emptyset$ and $S$ belong to $\mathcal{O}$. Further, if $\mathcal{C}$ is a subcollection of $\mathcal{O}$, then $S - \bigcup \mathcal{C} = \bigcap\{(S - L) \mid L \in \mathcal{C}\}$, which is a finite set because it is a subset of every finite set $S - L$, where $L \in \mathcal{C}$.

Also, if $U, V \in \mathcal{O}$, then $S - (U \cap V) = (S - U) \cup (S - V)$, which shows that $S - (U \cap V)$ is a finite set. Thus, $U \cap V \in \mathcal{O}$.

*Example 6.12.* Let $(S, \leq)$ be a partially ordered set. A subset $T$ of $S$ is *upward closed* if $x \in T$ and $x \leq y$ implies $y \in T$. The collection of upwards closed sets $\mathcal{O}^{\uparrow}$ is a topology on $S$.

It is clear that both $\emptyset$ and $S$ belong to $\mathcal{O}^{\uparrow}$. Further, if $\{L_i \mid i \in I\}$ is a family of upwards closed sets, then $\bigcup\{L_i \mid i \in I\}$ is also an upwards closed set. Indeed, suppose that $x \in \bigcup\{L_i \mid i \in I\}$ and $x \leq y$. There exists $L_i$ such that $x \in L_i$ and therefore $y \in L_i$, which implies $y \in \bigcup\{L_i \mid i \in I\}$. Moreover, it is easy to see that any intersection of sets from $\mathcal{O}^{\uparrow}$ belongs to $\mathcal{O}^{\uparrow}$, not just a finite intersection (which would suffice for $\mathcal{O}^{\uparrow}$ to be a topology). This topology is known as the *Alexandrov topology* on the poset $(S, \leq)$.

**Definition 6.13.** *A topology $\mathcal{O}$ is* finer *than a topology $\mathcal{O}'$ or, equivalently, $\mathcal{O}'$ is a* coarser *than $\mathcal{O}$, if $\mathcal{O}' \subseteq \mathcal{O}$.*

Every topology on a set $S$ is finer than the indiscrete topology on $S$; the discrete topology $\mathcal{P}(S)$ (which has the largest collection of open sets) is finer than any topology on $S$.

**Theorem 6.14.** *Let $(S, \mathcal{O})$ be a topological space and let $T$ be a subset of $S$. The collection $\mathcal{O} \restriction_T$ defined by*

$$\mathcal{O} \restriction_T = \{L \cap T \mid L \in \mathcal{O}\}$$

*is a topology on the set $T$.*

*Proof.* We leave the proof of this theorem to the reader as an exercise. □

**Definition 6.15.** *If $U$ is a subset of $S$, where $(S, \mathcal{O})$ is a topological space, then we refer to the topological space $(U, \mathcal{O} \restriction_U)$ as a* subspace *of the topological space $(S, \mathcal{O})$.*

To simplify notation, we shall denote the subspace $(U, \mathcal{O} \restriction_U)$ just by $U$.

**Theorem 6.16.** *Let $(S, \mathcal{O})$ be a topological space and let $(T, \mathcal{O} \restriction_T)$ be a subspace of this space. Then, a set $H$ is closed in $(T, \mathcal{O} \restriction_T)$ if and only if there exists a closed set $H_0$ in $(S, \mathcal{O})$ such that $H = T \cap H_0$.*

*Proof.* Suppose that $H$ is closed in $(T, \mathcal{O} \restriction_T)$. Then, the set $T - H$ is open in this space and therefore there exists an open set $L_0$ in $(S, \mathcal{O})$ such that $T - H = T \cap L_0$. This is equivalent to $H = T - (T \cap L_0) = T \cap (S - L_0)$. We define $H_0$ as the closed set $S - L_0$.

Conversely, suppose that $H = T \cap H_0$, where $H_0$ is a closed set in $S$. Since $T - H = T \cap (S - H_0)$ and $S - H_0$ is an open set in $(S, \mathcal{O})$, it follows that $T - H$ is open in the subspace and therefore $H$ is closed. □

**Corollary 6.17.** *Let $(S, \mathcal{O})$ be a topological space and let $T \subseteq S$. Denote by $\boldsymbol{K}_S$ and $\boldsymbol{K}_T$ the closure operators of $(S, \mathcal{O})$ and $(T, \mathcal{O} \restriction_T)$, respectively. For every subset $W$ of $T$, we have $\boldsymbol{K}_T(W) = \boldsymbol{K}_S(W) \cap T$.*

*Proof.* The set $\mathbf{K}_S(W)$ is closed in $S$, so $\mathbf{K}_S(W) \cap T$ is closed in $T$ by Theorem 6.16. Since $W \subseteq \mathbf{K}_S(W) \cap T$, it follows that $\mathbf{K}_T(W) \subseteq \mathbf{K}_S(W) \cap T$.

To prove the converse inclusion, observe that we can write $\mathbf{K}_T(W) = T \cap H$, where $H$ is a closed set in $S$ because $\mathbf{K}_T(W)$ is a closed set in $T$. Since $W \subseteq H$, it follows that $\mathbf{K}_S(W) \subseteq H$, so $\mathbf{K}_S(W) \cap T \subseteq H \cap T = \mathbf{K}_T(W)$. □

**Corollary 6.18.** *Let $(S, \mathcal{O})$ be a topological space and let $T \subseteq S$. If $U$ is a subset of $S$, then*

$$\boldsymbol{K}_T(U \cap T) \subseteq \boldsymbol{K}_S(U) \cap T.$$

*Proof.* By applying Corollary 6.17 to the subset $U \cap T$ of $T$ we have

$$\mathbf{K}_T(U \cap T) = \mathbf{K}_S(U \cap T) \cap T.$$

The needed inclusion follows from the monotonicity of $\mathbf{K}_S$. □

**Definition 6.19.** *A set $U$ is* dense *in a topological space $(S, \mathcal{O})$ if $\boldsymbol{K}(U) = S$. A topological space is* separable *if there exists a countable set $U$ that is dense in $(S, \mathcal{O})$.*

**Theorem 6.20.** *If $T$ is a subspace of a separable topological space $(S, \mathcal{O})$, then $T$ itself is separable.*

*Proof.* Since $S$ is separable, there exists a countable set $U$ such that $\mathbf{K}_S(U) = S$. On the other hand, $\mathbf{K}_T(U \cap T) = \mathbf{K}_S(U \cap T) \cap T \subseteq \mathbf{K}_S(U) \cap T = S \cap T = T$, which implies that the countable set $U \cap T$ is dense in $T$. Thus, $T$ is separable. □

**Theorem 6.21.** *If $T$ is a separable subspace of a topological space $(S, \mathcal{O})$, then so is $\boldsymbol{K}_S(T)$.*

*Proof.* Let $U$ be a countable subset of $T$ that is dense in $T$, that is, $\mathbf{K}_T(U) = T$. We need to prove that $\mathbf{K}_{\mathbf{K}_S(T)}(U) = \mathbf{K}_S(T)$ to prove that $U$ is dense in $\mathbf{K}_S(T)$ also.

By Corollary 6.17, we have

$$\mathbf{K}_{\mathbf{K}_S(T)}(U) = \mathbf{K}_S(U) \cap \mathbf{K}_S(T) = \mathbf{K}_S(U)$$

due to the monotonicity of $\mathbf{K}_S$.

Note that $T = \mathbf{K}_T(U) = \mathbf{K}_S(U) \cap T$, so $T \subseteq \mathbf{K}_S(U)$, which implies $\mathbf{K}_S(T) \subseteq \mathbf{K}_S(U)$. Since $\mathbf{K}_S$ is monotonic, we have the reverse inclusion $\mathbf{K}_S(U) \subseteq \mathbf{K}_S(T)$, so $\mathbf{K}_S(U) = \mathbf{K}_S(T)$. This allows us to conclude that $\mathbf{K}_{\mathbf{K}_S(T)}(U) = \mathbf{K}_S(T)$, so $U$ is dense in $\mathbf{K}_S(T)$. □

**Theorem 6.22.** *Let $(S, \mathcal{O})$ be a topological space. The set $U$ is dense in $(S, \mathcal{O})$ if and only if $U \cap L \neq \emptyset$ for every non-empty open set $L$.*

*Proof.* Suppose that $U$ is dense, so $\mathbf{K}(U) = S$. Since $\mathbf{K}(U \cap L) = \mathbf{K}(U) \cap \mathbf{K}(L) = S \cap \mathbf{K}(L) = \mathbf{K}(L)$, $U \cap L = \emptyset$ would imply $\mathbf{K}(L) = \mathbf{K}(\emptyset) = \emptyset$, which is a contradiction because $\emptyset \neq L \subseteq \mathbf{K}(L)$.

Conversely, suppose that $U$ has a non-empty intersection with every non-empty open set $L$. Since $\mathbf{K}(U)$ is closed, $S - \mathbf{K}(U)$ is open. Observe that $U \cap (S - \mathbf{K}(U)) = \emptyset$, so the open set $S - \mathbf{K}(U)$ must be empty. Therefore, we have $\mathbf{K}(U) = S$. □

**Theorem 6.23.** *Let $(S, \mathcal{O})$ be a topological space and let $x \in S$. The following statements hold:*

(i) *If $U, V \in neigh_x(\mathcal{O})$, then $U \cap V \in neigh_x(\mathcal{O})$.*
(ii) *If $U \in neigh_x(\mathcal{O})$ and $U \subseteq W \subseteq S$, then $W \in neigh_x(\mathcal{O})$.*
(iii) *A set $L$ is open if and only if $L$ is a neighborhood of all its points.*

*Proof.* The first two parts follow immediately from Definition 6.6. We discuss here only the third statement.

If $L$ is open, it is immediate that $L$ is a neighborhood of all its points. Conversely, suppose that $L$ is a neighborhood of all its members. Then, for each $x \in L$ there exists $W_x \in \mathcal{O}$ such that $x \in W_x \subseteq L$. Therefore,

$$L = \bigcup_{x \in L} \{x\} \subseteq \bigcup_{x \in L} W_x \subseteq L,$$

which implies $L = \bigcup_{x \in L} W_x$. This in turn implies $L \in \mathcal{O}$. □

**Definition 6.24.** *Let $(S, \mathcal{O})$ be a topological space. A subset $U$ of $S$ is* clopen *if it is both open and closed.*

Clearly, in every topological space $(S, \mathcal{O})$, both $\emptyset$ and $S$ are clopen sets.

In Chapter 4, we discussed the notion of an interior system of sets on a set $S$ and the notion of an interior operator. Since $\emptyset$ is an open set in any topological

space $(S, \mathcal{O})$ and any union of open sets is an open set, it follows that the topology itself is an interior system on $S$. In addition, an interior system of open sets is closed to finite intersection. Definition 6.25 which follows is a restatement of the definition of the interior operator associated to an interior system contained by Theorem 4.50.

**Definition 6.25.** *Let $(S, \mathcal{O})$ be a topological space. The* interior *of a set $U$, $U \subseteq S$, is the set*

$$\boldsymbol{I}(U) = \bigcup \{L \in \mathcal{O} \mid L \subseteq U\}.$$

The interior $\mathbf{I}(U)$ of a set $U$ is the largest open set included in $U$, because the union of any collection of open sets is an open set. Furthermore, a set is open in a topological space if and only if it equals its interior.

**Theorem 6.26.** *Let $(S, \mathcal{O})$ be a topological space and let $U$ be a subset of $S$. The closure $\boldsymbol{K}(S - U)$ of the set $S - U$ equals $S - \boldsymbol{I}(U)$.*

*Proof.* Since $\mathbf{I}(U)$ is an open set, the set $S - \mathbf{I}(U)$ is closed. Note that $S - U \subseteq S - \mathbf{I}(U)$. Therefore, $\mathbf{K}(S - U) \subseteq S - \mathbf{I}(U)$.

Conversely, the inclusion $S - U \subseteq \mathbf{K}(S - U)$ implies $S - \mathbf{K}(S - U) \subseteq U$. Since $S - \mathbf{K}(S - U)$ is an open set included in $U$ and $\mathbf{I}(U)$ is the largest such set, it follows that $S - \mathbf{K}(S - U) \subseteq \mathbf{I}(U)$, which implies $S - \mathbf{I}(U) \subseteq \mathbf{K}(S - U)$. □

**Corollary 6.27.** *For every subset $U$ of a topological space $(S, \mathcal{O})$, we have*

$$\boldsymbol{I}(U) = S - \boldsymbol{K}(S - U)$$

*and*

$$\boldsymbol{K}(U) = S - \boldsymbol{I}(S - U).$$

*Proof.* The first equality is immediate; the second follows from Theorem 6.26 by replacing $U$ by $S - U$. □

**Theorem 6.28.** *Let $(S, \mathcal{O})$ be a topological space. The following statements are equivalent:*

(i) *Every countable intersection of dense open sets is a dense set.*
(ii) *Every countable union of closed sets that have an empty interior has an empty interior.*

*Proof.* (i) implies (ii): Let $H_1, \ldots, H_n, \ldots$ be a sequence of closed sets with $\mathbf{I}(H_i) = \emptyset$ for $n \geq 1$. Then, for the open sets $L_i$ given by $L_i = S - H_i$, we have $\mathbf{K}(L_i) = \mathbf{K}(S - H_i) = S - \mathbf{I}(H_i) = S$, so every set $L_i$ is dense. By (i), we have $\mathbf{K}(\bigcap_{i \geq 1} L_i) = S$, so

$$\begin{aligned}\mathbf{I}\left(\bigcup_{i\geq 1} H_i\right) &= S - \mathbf{K}\left(S - \bigcup_{i\geq 1} H_i\right)\\ &= S - \mathbf{K}\left(\bigcap_{i\geq 1}(S - H_i)\right)\\ &= S - \mathbf{K}\left(\bigcap_{i\geq 1} L_i\right) = \emptyset,\end{aligned}$$

which shows that (ii) holds.

(ii) implies (i): this argument is similar to the preceding one and we will omit it. □

A topological space that satisfies one of the equivalent conditions of this theorem is called a *Baire space*. As we shall see in Chapter 11 (Theorem 11.54), a very important category of topological spaces, the complete topological metric spaces, are Baire spaces.

**Definition 6.29.** *Let $(S, \mathcal{O})$ be a topological space. The* border *of a set $U$, where $U \in \mathcal{P}(S)$, is the set $\partial_S K = \boldsymbol{K}(U) \cap \boldsymbol{K}(S - U)$.*

*If $S$ is clear from the context, then we will omit the subscript and will denote the border of $U$ just by $\partial U$.*

The border itself is obviously a closed set, as it is an intersection of two closed sets.

Note that, by using Corollary 6.27, the border of a set can be expressed also in term of interiors:

$$\partial U = (S - \mathbf{I}(S - U)) \cap (S - \mathbf{I}(U)) = S - (\mathbf{I}(S - U) \cup \mathbf{I}(U)). \tag{6.2}$$

**Theorem 6.30.** *The border of a subset $U$ of a topological space $(S, \mathcal{O})$ consists of those elements $s$ of $S$ such that for every open set $L$ that contains $s$ we have both $L \cap U \neq \emptyset$ and $L \cap (S - U) \neq \emptyset$.*

*Proof.* Let $x \in \partial U$ and let $L$ be an open set such that $x \in L$. By Equality( 6.2), we have both $x \notin \mathbf{I}(S - U)$ and $x \notin \mathbf{I}(U)$. Therefore, $L \not\subseteq S - U$ and $L \not\subseteq U$, which imply $L \cap U \neq \emptyset$ and $L \cap (S - U) \neq \emptyset$.

Conversely, suppose that, for every open set $L$ that contains $s$, we have both $L \cap U \neq \emptyset$ and $L \cap (S - U) \neq \emptyset$. This implies $x \notin \mathbf{I}(U)$ and $s \notin \mathbf{I}(S - U)$, so $x \in \partial U$ by Equality (6.2). □

**Theorem 6.31.** *Let $(S, \mathcal{O})$ be a topological space, $(T, \mathcal{O} \restriction_T)$ be a subspace, and $W$ be a subset of $S$. The border $\partial_T(W \cap T)$ of $W \cap T$ in the subspace $T$ is a subset of the intersection $\partial_S(W) \cap T$, where $\partial_S(W)$ is the border of $W$ in $S$.*

*Proof.* By Definition 6.29, we have

$$\begin{aligned}\partial_T(W \cap T) &= \mathbf{K}_T(W \cap T) \cap \mathbf{K}_T(T - (W \cap T)) \\ &= \mathbf{K}_T(W \cap T) \cap \mathbf{K}_T(T - W) \\ &\subseteq (\mathbf{K}_S(W) \cap T) \cap \mathbf{K}_T(T - W) \\ &\quad \text{by Corollary 6.17.}\end{aligned}$$

Again, by Corollary 6.17, we have $\mathbf{K}_T(T-W) = \mathbf{K}_T(T \cap (S-W)) \subseteq \mathbf{K}_S(S-W) \cap T$, and this allows us to write

$$\partial_T(W \cap T) \subseteq (\mathbf{K}_S(W) \cap T) \cap \mathbf{K}_S(S - W) \cap T = \partial_S(W) \cap T,$$

which is the desired conclusion. □

The next statement relates three important sets that we defined for each subset $U$ of a topological space $(S, \mathcal{O})$.

**Theorem 6.32.** *Let $(S, \mathcal{O})$ be a topological space. For every subset $U$ of $S$, we have $\boldsymbol{K}(U) = \boldsymbol{I}(U) \cup \partial U$.*

*Proof.* By Equality (6.2), we have $\partial U = (S-\mathbf{I}(S-U)) \cap (S-\mathbf{I}(U))$. Therefore,

$$\begin{aligned}\partial U \cup \mathbf{I}(U) &= (S - \mathbf{I}(S - U)) \cap \mathbf{I}(U) \\ &\quad \text{(by Corollary 6.27)} \\ &= \mathbf{K}(U) \cap \mathbf{I}(U) \\ &\quad \text{(because } \mathbf{I}(U) \subseteq \mathbf{K}(U)\text{)} \\ &= \mathbf{K}(U).\end{aligned}$$

□

**Corollary 6.33.** *Let $(S, \mathcal{O})$ be a topological space and let $(T, \mathcal{O} \restriction_T)$ be a subspace of $(S, \mathcal{O})$. For any subset $U$ of $S$, we have $\partial_T(U \cap T) \subseteq \partial_S(U)$.*

*Proof.* Let $t \in \partial_T(U \cap T)$. By Theorem 6.30, for every open set $L \in \mathcal{O} \restriction_T$ such that $t \in L$ we have both $L \cap (U \cap T) \neq \emptyset$ and $L \cap (T - (U \cap T)) \neq \emptyset$.

If $L_1$ is an open set of $(S, \mathcal{O})$ that contains $S$, then $L_1 \cap T$ is an open set of $(T, \mathcal{O} \restriction_T)$ that contains $t$, so for $L_1$ we have both $(L_1 \cap T) \cap (U \cap T) \neq \emptyset$ and $(L_1 \cap T) \cap (T - (U \cap T)) \neq \emptyset$. This immediately implies $L_1 \cap U \neq \emptyset$ and $L_1 \cap (S - U) \neq \emptyset$, that is, $t \in \partial_S(U)$. □

**Theorem 6.34.** *Let $(S, \mathcal{O})$ be a topological space. A set $U$ is clopen if and only if $\partial U = \emptyset$.*

*Proof.* Suppose that $U$ is clopen. Then $U = \mathbf{K}(U)$; moreover, $S - U$ is also closed (because $U$ is open) and therefore $S - U = \mathbf{K}(S - U)$. Thus, $\mathbf{K}(U) \cap \mathbf{K}(S - U) = U \cap (S - U) = \emptyset$, so $\partial U = \emptyset$.

Conversely, suppose that $\partial U = \emptyset$. Then, since $\mathbf{K}(U) \cap \mathbf{K}(S - U) = \emptyset$, it follows that $\mathbf{K}(U) \subseteq S - \mathbf{K}(S - U)$. Therefore, $\mathbf{K}(U) \subseteq S - (S - U) = U$, which implies $\mathbf{K}(U) = U$. Thus, $U$ is closed. Furthermore, by Equality (6.2), $\partial U = \emptyset$ also implies $\mathbf{I}(S - L) \cup \mathbf{I}(L) = S$, so $S - \mathbf{I}(S - L) \subseteq \mathbf{I}(L)$. By Corollary 6.27, we have $\mathbf{K}(L) \subseteq \mathbf{I}(L)$, so $L \subseteq \mathbf{I}(L)$. Thus, $L = \mathbf{I}(L)$, so $L$ is also an open set. □

**Definition 6.35.** *Let $(S, \mathcal{O})$ be a topological space and let $U$ be a subset of $S$. An element $t$ of $S$ is an* accumulation point *or a* cluster point *of the set $U$ if, for every open set $L$ such that $t \in L$, the set $U \cap (L - \{t\})$ is not empty.*

*The set of all accumulation points of a set $U$ is the* derived set *of $U$ and is denoted by $U'$.*

**Lemma 6.36.** *Let $(S, \mathcal{O})$ be a topological space and let $U$ be a subset of $S$. We have $\partial U \subseteq U \cup U'$.*

*Proof.* By Theorem 6.30, if $x \in \partial U$, then for every open set $L$ such that $x \in L$, we have both $L \cap U \neq \emptyset$ and $L \cap (S - U) \neq \emptyset$.

If $U \cap (L - \{x\}) \neq \emptyset$ for every open set $L$, then $x \in U'$. Otherwise, there is an open set $L_0$ such that $x_0 \in L$ and $U \cap (L_0 - \{x\}) = \emptyset$. This can happen only if $x \in U$. Therefore, in either case, $x \in U \cup U'$, which gives the desired inclusion. □

**Theorem 6.37.** *Let $(S, \mathcal{O})$ be a topological space and let $U$ be a subset of $S$. We have $\boldsymbol{K}(U) = U \cup U'$ for every subset $U$ of $S$.*

*Proof.* By Theorem 6.32 and Lemma 6.36, we have $\mathbf{K}(U) = \mathbf{I}(U) \cup \partial U \subseteq \mathbf{I}(U) \cup U \cup U' = U \cup U'$ because $\mathbf{I}(U) \subseteq U$.

Let $x$ be an accumulation point of $U$. If $x \in U$, then clearly $x \in \mathbf{K}(U)$. Otherwise, $x \notin U$ and we claim that in this case $x \in \mathbf{K}(U)$. Indeed, if $x$ were not an element of $\mathbf{K}(U)$, it would belong to the open set $S - \mathbf{K}(U)$. This would imply that the set $U \cap (S - \mathbf{K}(U) - \{x\})$ is not empty, which is a contradiction. This yields the reverse inclusion, $U \cup U' \subseteq \mathbf{K}(U)$. □

## 6.4 Bases

Let $\mathfrak{O} = \{\mathcal{O}_i \mid i \in I\}$ be a family of topologies defined on a set $S$ that contains the discrete topology $\mathcal{P}(S)$. We claim that $\mathfrak{O}$ is a closure system on $\mathcal{P}(S)$. The first condition of Definition 4.32 is satisfied due to the definition of $\mathfrak{O}$. It is easy to verify that for every subfamily $\mathfrak{O}'$ of $\mathfrak{O}$, $\bigcap \mathfrak{O}'$ is a topology, so $\mathfrak{O}$ is indeed a closure system.

Thus, if $\mathcal{S}$ is a family of subsets of $S$, there exists the smallest topology that includes $\mathcal{S}$.

**Theorem 6.38.** *The topology $TOP(\mathcal{S})$ generated by a family $\mathcal{S}$ of subsets of $S$ consists of unions of finite intersections of the members of $\mathcal{S}$.*

*Proof.* Let $\mathcal{E}$ be the collection of all unions of finite intersections of the members of $\mathcal{S}$. It is clear that $\mathcal{S} \subseteq \mathcal{E}$. We claim that $\mathcal{E}$ is a topology that contains $\mathcal{S}$.

Note that the intersection of the empty collection of sets in $\mathcal{S}$ is $S$, so $S \in \mathcal{E}$; also, the union of an empty collection of finite intersections is $\emptyset$, so $\emptyset \in \mathcal{E}$.

Every $U \in \mathcal{E}$ can be written as

$$U = \bigcup \{V_j \mid j \in J_U\},$$

where the sets $V_j$ are finite intersections of sets of $\mathcal{S}$. Therefore, it is immediate that any union of sets of this form belongs to $\mathcal{E}$.

Suppose that $\{U_i \mid i \in I\}$ is a finite collection of parts of $S$, where $U_i = \bigcup \{V_j \in \mathcal{S} \mid j \in J_i\}$ and that each $V_j$ can be written as $V_j = \bigcap \{W_{jh} \in \mathcal{S} \mid h \in H_j\}$, where each set $H_j$ is finite. One can prove by induction on $p = |I|$ that $\bigcap \{U_i \mid i \in I\} \in \mathcal{E}$. To simplify the presentation, we discuss here only the case where $|I| = 2$. So, if $U_i = \bigcup \{V_j \in \mathcal{S} \mid j \in J_i\}$ for $i = 1, 2$, we have

$$\begin{aligned} U_1 \cap U_2 &= \bigcup \{V_{j_1} \in \mathcal{S} \mid j_1 \in J_1\} \cap \bigcup \{V_{j_2} \in \mathcal{S} \mid j_2 \in J_2\} \\ &= \bigcap_{j_1, j_2} (V_{j_1} \cap V_{j_2}). \end{aligned}$$

Since each intersection $V_{j_1} \cap V_{j_2}$ is in turn a finite intersection of sets of $\mathcal{S}$, it follows that $U_1 \cap U_2 \in \mathcal{S}$.

Thus, $TOP(\mathcal{S})$ is contained in $\mathcal{E}$ because $TOP(\mathcal{S})$ is the coarsest topology that contains $\mathcal{S}$. This gives the desired conclusion. □

**Corollary 6.39.** *Let $\mathcal{B}$ be a collection of subsets of the set $S$ such that for every finite subcollection $\mathcal{D}$ of $\mathcal{B}$, $x \in \bigcap \mathcal{D}$ implies the existence of a set $B \in \mathcal{B}$ such that $x \in B \subseteq \bigcap \mathcal{D}$. Then, $TOP(\mathcal{B})$, the topology generated by $\mathcal{B}$, consists of sets that are unions of subcollections of $\mathcal{B}$.*

*Proof.* By Theorem 6.38, $TOP(\mathcal{B})$ consists of unions of finite intersections of the members of $\mathcal{B}$. Therefore, unions of sets of $\mathcal{B}$ belong to $TOP(\mathcal{B})$.

Conversely, let $U \in \mathcal{B}$, that is, $U = \bigcup \{V_i \mid i \in I\}$, where each $V_i$ is a finite intersection of members of $\mathcal{B}$. For every $x \in V_i$, there exists a set $B_{x,i} \in \mathcal{B}$ such that $x \in B_{x,i} \subseteq V_i$. Therefore, $V_i = \bigcup_{x \in V_i} B_{x,i}$, and this implies that $U$ is indeed a union of sets from $\mathcal{B}$. □

**Definition 6.40.** *Let $S$ be a set. A collection $\mathcal{S}$ is a* subbasis *for a topology $\mathcal{O}$ if $\mathcal{O} = TOP(\mathcal{S})$.*

*A collection $\mathcal{B}$ of subsets is a* basis *for a topology if, for every finite subcollection $\mathcal{D}$ of $\mathcal{B}$, if $x \in \bigcap \mathcal{D}$, then there exists a set $B \in \mathcal{B}$ such that $x \in B \subseteq \bigcap \mathcal{D}$.*

Corollary 6.39 implies that, for a basis $\mathcal{B}$, we have $\bigcup \mathcal{B} = S$. Indeed, consider the intersection of the empty collection of parts of $\mathcal{B}$, which equals $S$. Then, for every $x \in S$, there is a set $B \in \mathcal{B}$ such that $x \in B \subseteq S$, which of course implies $\bigcup \mathcal{B} = S$.

Clearly, every set of $\mathcal{B}$ is an open set in the topological space $(S, TOP(\mathcal{B})$.

Starting from a topology, we find a basis using the following theorem.

**Theorem 6.41.** *Let $(S, \mathcal{O})$ be a topological space. If $\mathcal{B}$ is a collection of open subsets of $S$ such that for every $x \in S$ and every open set $L \in \mathcal{O}$ there exists a set $B \in \mathcal{B}$ such that $x \in B \subseteq L$, then $\mathcal{B}$ is a basis for $(S, \mathcal{O})$.*

*Proof.* This statement is an immediate consequence of Definition 6.40. □

**Theorem 6.42.** *Let $(S, \mathcal{O})$ be a topological space. The following statements involving a family $\mathcal{B}$ of subsets of $S$ are equivalent.*

(i) *$\mathcal{B}$ is a basis for $(S, \mathcal{O})$;*

(ii) *For every $x \in S$ and $U \in neigh_x(\mathcal{O})$, there exists $B \in \mathcal{B}$ such that $x \in B \subseteq U$.*

(iii) *For every open set $L$, there is a subcollection $\mathcal{C}$ of $\mathcal{B}$ such that $L = \bigcup \mathcal{C}$.*

*Proof.* (i) implies (ii): Let $\mathcal{B}$ be a basis for $(S, \mathcal{O})$ and let $U \in neigh_x(\mathcal{O})$. There exists an open set $L$ such that $x \in L \subseteq U$. Since $\mathcal{B}$ is a basis, there exists a set $B \in \mathcal{B}$ such that $x \in B \subseteq L \subseteq U$, which is what we aimed to prove.

(ii) implies (iii): Suppose that the second statement holds, and let $L$ be an open set. Since $L$ is a neighborhood for all its elements, for every $x \in L$ there exists $B_x \in \mathcal{B}$ such that $\{x\} \subseteq B_x \subseteq L$. Therefore, $L = \bigcup \{B_x \mid x \in L\}$.

(iii) implies (i): Part (iii) implies Part (i) immediately. □

**Corollary 6.43.** *Let $U$ be a subspace of a topological space $(S, \mathcal{O})$. If $\mathcal{B}$ is a basis of $(S, \mathcal{O})$, then $\mathcal{B}_U = \{U \cap B \mid B \in \mathcal{B}\}$ is a basis of the subspace $U$.*

*Proof.* Let $K$ be an open subset in the subspace $U$. There is an open set $L$ in $(S, \mathcal{O})$ such that $K = U \cap L$. Since $\mathcal{B}$ is a basis for $(S, \mathcal{O})$, by the third part of Theorem 6.42, there is a subcollection $\mathcal{C}$ of $\mathcal{B}$ such that $L = \bigcup \mathcal{C}$, which implies $K = \bigcup \{U \cap C \mid C \in \mathcal{C}\}$. Thus, $\mathcal{B}_U$ is a basis for $U$. □

*Example 6.44.* The collection of open intervals $\{(a, b) \mid a, b \in \mathbb{R} \text{ and } a < b\}$ is a basis for the topological space $(\mathbb{R}, \mathcal{O})$ by Theorem 6.10. Further, the collection

$$\mathcal{S} = \{(a, +\infty) \mid a \in \mathbb{R}\} \cup \{(-\infty, b) \mid b \in \mathbb{R}\}$$

is a subbasis of this topology because every member $(a, b)$ of the basis can be written as $(a, b) = (-\infty, b) \cap (a, +\infty)$.

**Definition 6.45.** *A topological space satisfies the* first axiom of countability *if for every $x \in S$ there is a countable family of open sets $\mathcal{L}_x = \{L_n \mid n \in \mathbb{N}\}$ such that $x \in \bigcap\{L_n \mid n \in \mathbb{N}\}$ and for every open set $L$ that contains $x$ there is a set $L_n \in \mathcal{L}_x$ such that $L_n \subseteq L$.*

*A topological space satisfies the* second axiom of countability *if it has a countable basis.*

It is clear that the second axiom of countability implies the first, and we will deal mostly with this second axiom. Furthermore, by Corollary 6.43, every subspace of a topological space that satisfies the second axiom of countability satisfies this axiom itself.

**Theorem 6.46.** *Let $(S, \mathcal{O})$ be a topological space. If $(S, \mathcal{O})$ has a countable basis, then $(S, \mathcal{O})$ is separable.*

*Proof.* Let $\{B_n \mid n \in \mathbb{N}\}$ be a countable basis for $(S, \mathcal{O})$ and let $x_n$ be an element of $B_n$ for $n \in \mathbb{N}$. We claim that $S = \mathbf{K}(\{x_n \mid n \in \mathbb{N}\})$, which is equivalent to $S - \mathbf{K}(\{x_n \mid n \in \mathbb{N}\}) = \emptyset$.

Indeed, observe that $S - \mathbf{K}(\{x_n \mid n \in \mathbb{N}\})$ is a non-empty open set; therefore, there exists $m \in \mathbb{N}$ such that $B_m \subseteq S - \mathbf{K}(\{x_n \mid n \in \mathbb{N}\})$, so $x_m \in S - \mathbf{K}(\{x_n \mid n \in \mathbb{N}\}) \subseteq S - \{x_n \mid n \in \mathbb{N}\}$, which is a contradiction. Therefore, the countable set $\{x_n \mid n \in \mathbb{N}\}$ is dense in $(S, \mathcal{O})$. □

The notion of an open cover of a topological space is introduced next.

**Definition 6.47.** *A* cover *of a topological space $(S, \mathcal{O})$ is a collection of sets $\mathcal{C}$ such that $\bigcup \mathcal{C} = S$.*

*If $\mathcal{C}$ is a cover of $(S, \mathcal{O})$ and every set $C \in \mathcal{C}$ is open (closed), then we refer to $\mathcal{C}$ as an* open cover *(a* closed cover*, respectively).*

*A* subcover of an open cover $\mathcal{C}$ *is a collection $\mathcal{D}$ such that $\mathcal{D} \subseteq \mathcal{C}$ and $\bigcup \mathcal{D} = S$.*

**Theorem 6.48.** *If a topological space $(S, \mathcal{O})$ satisfies the second axiom of countability, then every basis $\mathcal{B}$ for $(S, \mathcal{O})$ contains a countable collection $\mathcal{B}_0$ that is a basis for $(S, \mathcal{O})$.*

*Proof.* Let $\mathcal{B}' = \{L_i \mid i \in \mathbb{N}\}$ be a countable basis for $(S, \mathcal{O})$ and let $\mathcal{C}_i$ be the subcollection of $\mathcal{B}$ defined by $\mathcal{C}_i = \{V \in \mathcal{B} \mid V \subseteq L_i\}$ for $i \in \mathbb{N}$. Since $\mathcal{B}$ is a basis for $(S, \mathcal{O})$, it is clear that $\mathcal{C}_i$ is an open cover for $L_i$; that is, $\bigcup \mathcal{C}_i = L_i$ for every $i \in \mathbb{N}$. Since each subspace $L_i$ has a countable basis, $\mathcal{C}_i$ contains a countable subcover $\mathcal{C}'_i$ of $L_i$. The collection $\mathcal{B}_0 = \bigcup\{\mathcal{C}'_i \mid i \in \mathbb{N}\}$ is countable and is a basis for $(S, \mathcal{O})$ that is included in $\mathcal{B}$. □

**Corollary 6.49.** *If a topological space $(S, \mathcal{O})$ has a countable basis, then every open cover of $(S, \mathcal{O})$ contains a countable subcover.*

*Proof.* This fact follows directly from Theorem 6.48. □

## 6.5 Compactness

**Definition 6.50.** *A topological space $(S, \mathcal{O})$ is* compact *if every open cover $\mathcal{C}$ of this space contains a finite subcover.*

Another useful concept is the notion of a family of sets with the finite intersection property.

**Definition 6.51.** *A collection $\mathcal{C}$ of subsets of a set $S$ has the* finite intersection property *(f.i.p.) if $\bigcap \mathcal{D} \neq \emptyset$ for every finite subcollection $\mathcal{D}$ of $\mathcal{C}$.*

**Theorem 6.52.** *The following three statements concerning a topological space $(S, \mathcal{O})$ are equivalent:*
(i) *$(S, \mathcal{O})$ is compact.*
(ii) *If $\mathcal{D}$ is a family of closed subsets of $S$ such that $\bigcap \mathcal{D} = \emptyset$, then there exists a finite subfamily $\mathcal{D}_0$ of $\mathcal{D}$ such that $\bigcap \mathcal{D}_0 = \emptyset$.*
(iii) *If $\mathcal{E}$ is a family of closed sets having the f.i.p., then $\bigcap \mathcal{E} \neq \emptyset$.*

*Proof.* The argument is left to the reader. □

Another characterization of compactness that is just a variant of Part (iii) of Theorem 6.52 that applies to an arbitrary family of sets (not necessarily closed) is given next.

**Theorem 6.53.** *A topological space $(S, \mathcal{O})$ is compact if and only if for every family of subsets $\mathcal{C}$ that has the f.i.p., $\bigcap\{\boldsymbol{K}(C) \mid C \in \mathcal{C}\} \neq \emptyset$.*

*Proof.* If for every family of subsets $\mathcal{C}$ that has the f.i.p. we have $\bigcap\{\mathbf{K}(C) \mid C \in \mathcal{C}\} \neq \emptyset$, then, in particular, if $\mathcal{C}$ consists of closed sets, it follows that $\bigcap\{C \mid C \in \mathcal{C}\} \neq \emptyset$, which amounts to Part (iii) of Theorem 6.52, so $(S, \mathcal{O})$ is compact.

Conversely, suppose that the space $(S, \mathcal{O})$ is compact, which means that the property of Part (iii) of Theorem 6.52 holds. Suppose that $\mathcal{C}$ is an arbitrary collection of subsets of $S$ that has the f.i.p. Then, the collection of closed subsets $\{\mathbf{K}(C) \mid C \in \mathcal{C}\}$ also has the f.i.p. because $C \in \mathbf{K}(\mathcal{C})$ for every $C \in \mathcal{C}$. Therefore, $\bigcap\{\mathbf{K}(C) \mid C \in \mathcal{C}\} \neq \emptyset$. □

*Example 6.54.* Let $U_1 \supseteq U_2 \supseteq \cdots$ be a descending sequence of non-empty closed subsets of a compact space $(S, \mathcal{O})$. Its intersection $\bigcap_{n \geq 1} U_n$ is non-empty because $(S, \mathcal{O})$ is compact and $\bigcap_{p=1}^{k} U_{i_p} = U_l \neq \emptyset$, where $l = \min\{i_1, \ldots, i_k\}$ for every $k \geq 1$.

This implies that the topological space $(\mathbb{R}, \mathcal{O})$ introduced in Example 6.4 is not compact because $\bigcap_{n \geq 1} [n, \infty) = \emptyset$.

The notion of cover refinement can be used to characterize compact topological spaces. Recall that we introduced this notion in Definition 1.12.

**Theorem 6.55.** *A topological space $(S, \mathcal{O})$ is compact if and only if every open cover $\mathcal{C}$ is refined by some finite open cover of the space.*

*Proof.* Suppose that $(S, \mathcal{O})$ is compact. Then, every open cover $\mathcal{C}$ contains a finite subcover $\mathcal{C}'$. Since every $C' \in \mathcal{C}'$ is a member of $\mathcal{C}$, it follows that $\mathcal{C}$ is refined by $\mathcal{C}'$.

Conversely, suppose that every open cover $\mathcal{C}$ is refined by some finite open cover $\mathcal{D} = \{D_1, \ldots, D_p\}$. Then, for every $D_i \in \mathcal{D}$ there exists a set $C_i \in \mathcal{C}$ such that $D_i \subseteq C_i$ for $1 \leq i \leq p$. Since $\bigcup_{i=1}^n D_i = S$, it follows that $\bigcup_{i=1}^n C_i = S$, so $\{C_1, \ldots, C_n\}$ is a finite subcover of $\mathcal{C}$, which means that $(S, \mathcal{O})$ is compact. □

If $(T, \mathcal{O} \restriction_T)$ is a compact topological space, then we say that $T$ is a *compact set.*

*Example 6.56.* Every closed interval $[x, y]$ of $\mathbb{R}$ is a compact set. Indeed, if $\mathcal{C}$ is an open cover of $[x, y]$ we can assume without loss of generality that $\mathcal{C}$ consists of open intervals $\mathcal{C} = \{(a_i, b_i) \mid i \in I\}$.

Let

$$K = \left\{ c \mid c \in [x, y] \text{ and } [x, c] \subseteq \bigcup_{j \in J} (a_j, b_j) \text{ for some finite } J \subseteq I \right\}.$$

Observe that $K \neq \emptyset$ because $x \in K$. Indeed, we have $[x, x] = \{x\}$ and therefore $[x, x] \subseteq (a_i, b_i)$ for some $i \in I$.

We claim that $y \leq w = \sup K$. It is clear that $w \leq y$ because $y$ is an upper bound of $[x, y]$ and therefore an upper bound of $K$. Suppose that $w < y$. Note that in this case there exists an open interval $(a_p, b_p)$ for some $p \in I$ such that $w \in (a_p, b_p)$. By Theorem 4.28, for every $\epsilon > 0$, there is $z \in K$ such that $\sup K - \epsilon < z$. Choose $\epsilon$ such that $\epsilon < w - a_p$. Since the closed interval $[x, z]$ is covered by a finite collection of open intervals $[x, z] \subseteq (a_{j_1}, b_{j_1}) \cup \cdots \cup (a_{j_r}, b_{j_r})$, it follows that the interval $[x, w]$ is covered by $(a_{j_1}, b_{j_1}) \cup \cdots \cup (a_{j_r}, b_{j_r}) \cup (a_p, b_p)$. This leads to a contradiction because the open interval $(a_p, b_p)$ contains numbers in $K$ that are greater than $w$. So we have $w = y$, which shows that $[x, y]$ can be covered by a finite family of open intervals extracted from $\mathcal{C}$.

*Example 6.57.* The open interval $(0, 1)$ is not compact. Indeed, it is easy to see that the collection of open sets $\left\{\left(\frac{1}{n}, 1 - \frac{1}{n}\right)\right\}$ is an open cover of $(0, 1)$. However, no finite sub-collection of this collection of sets is an open cover of $(0, 1)$.

Example 6.57 suggests the interest of the following definition.

**Definition 6.58.** *A subset $T$ of a topological space is* relatively compact *if its closure $\boldsymbol{K}(T)$ is compact.*

*Example 6.59.* The set $(0,1)$ is a relatively compact subset of $\mathbb{R}$ but not a compact one.

**Theorem 6.60.** *If $(S, \mathcal{O})$ is a compact topological space, any closed subset $T$ of $S$ is compact.*

*Proof.* Let $T$ be a closed subset of $(S, \mathcal{O})$. We need to show that the subspace $(T, \mathcal{O} \restriction_T)$ is compact. Let $\mathcal{C}$ be an open cover of the space $(T, \mathcal{O} \restriction_T)$. Then, $\mathcal{C} \cup \{S - T\}$ is a open cover of $(S, \mathcal{O})$. The compactness of $(S, \mathcal{O})$ means that there exists a finite subcover $\mathcal{D}$ of $(S, \mathcal{O})$ such that $\mathcal{D} \subseteq \mathcal{C} \cup \{S - T\}$. It follows immediately that $\mathcal{D} - \{S - T\}$ is a finite subcover of $\mathcal{C}$ for $(T, \mathcal{O} \restriction_T)$. □

A topological space $(S, \mathcal{O})$ is *locally compact* if for every $x \in S$ there exists an open set $L \in \mathcal{O}$ such that $x \in L$ and $\mathbf{K}(L)$ is a compact set.

**Theorem 6.61.** *If $(S, \mathcal{O})$ is a compact topological space, then, for every infinite subset $U$ of $S$ we have $U' \neq \emptyset$ (the* Bolzano-Weierstrass property*).*

*Proof.* Let $U = \{x_i \mid i \in I\}$ be an infinite subset of $S$. Suppose that $U$ has no accumulation point. For every $s \in S$, there is an open set $L_s$ such that $s \in L_s$ and $U \cap (L_s - \{s\}) = \emptyset$. Clearly the collection $\{L_s \mid s \in S\}$ is an open cover of $S$, so it contains a finite subcover $\{L_{s_1}, \ldots, L_{s_p}\}$. Thus, $S = L_{s_1} \cup \cdots \cup L_{s_p}$. Note that each $L_{s_i}$ contains at most one element of $U$ (which happens when $s_i \in U$), which implies that $U$ is finite. This contradiction means that $U' \neq \emptyset$. □

## 6.6 Continuous Functions

The notion of a continuous function is central in topology; we introduce it in the next definition.

**Definition 6.62.** *Let $(S_1, \mathcal{O}_1)$ and $(S_2, \mathcal{O}_2)$ be two topological spaces. A function $f : S_1 \longrightarrow S_2$ is* continuous *if, for every open set $V \in \mathcal{O}_2$, we have $f^{-1}(V) \in \mathcal{O}_1$.*

If $f : S_1 \longrightarrow S_2$ is a continuous function between the topological spaces $(S_1, \mathcal{O}_1)$ and $(S_2, \mathcal{O}_2)$ and $\mathcal{O}_1'$ and $\mathcal{O}_2'$ are topologies on $S_1$ and $S_2$, respectively, such that $\mathcal{O}_2' \subseteq \mathcal{O}_2$ and $\mathcal{O}_1 \subseteq \mathcal{O}_1'$, then $f$ is also a continuous function between the topological spaces $(S_1, \mathcal{O}_1')$ and $(S_2, \mathcal{O}_2')$. Therefore, any function defined on the topological space $(S, \mathcal{P}(S))$ (equipped with the discrete topology) with values in an arbitrary topological space $(S', \mathcal{O}')$ is continuous; similarly, any function $f : S \longrightarrow S'$ between a topological space $(S, \mathcal{O})$ and $(S', \{\emptyset, S'\})$ (equipped with the discrete topology) is continuous.

**Theorem 6.63.** *Let $(S, \mathcal{O}), (T, \mathcal{O}')$, and $(U, \mathcal{O}'')$ be three topological spaces and let $f : S \longrightarrow T$ and $g : T \longrightarrow U$ be two continuous functions. Then, the function $gf : S \longrightarrow T$ is continuous.*

*Proof.* This statement is an immediate consequence of Definition 6.62 and Theorem 1.63. □

The next theorem provides several equivalent characterizations of continuous functions and thus gives several alternative methods for proving the continuity of a function.

**Theorem 6.64.** *Let $(S, \mathcal{O})$ and $(T, \mathcal{O}')$ be two topological spaces and let $f : S \longrightarrow T$ be a function. The following statements are equivalent:*

(i) *$f$ is continuous.*
(ii) *For every closed set $L$, $L \subseteq T$, the set $f^{-1}(L)$ is a closed set in $(S, \mathcal{O})$.*
(iii) *$f(\boldsymbol{K}_1(H)) \subseteq \boldsymbol{K}_2(f(H))$ for every $H \subseteq S$, where $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are the closure operators of the topological spaces $(S, \mathcal{O})$ and $(T, \mathcal{O}')$, respectively.*
(iv) *For every $x \in S$ and $V \in neigh_{f(x)}(\mathcal{O}')$, there exists $U \in neigh_x(\mathcal{O})$ such that $f(U) \subseteq V$.*

*Proof.* To prove that (i) implies (ii), let $f$ be a continuous function and let $C$ be the open set given by $C = T - L$. By (i), $f^{-1}(C)$ is open in $(S, \mathcal{O})$ and therefore $S - f^{-1}(C)$ is closed in $(S, \mathcal{O})$. Since

$$S - f^{-1}(C) = S - f^{-1}(T - L) = f^{-1}(L)$$

(see Exercise 21), we have shown the desired implication.

To prove that (ii) implies (iii), we start from the fact that $H \subseteq f^{-1}(f(H))$. Therefore, $H \subseteq f^{-1}(\mathbf{K}_2(f(H)))$. Since $\mathbf{K}_2(f(H))$ is closed, it follows that $f^{-1}(\mathbf{K}_2(f(H)))$ is also closed. Thus, $\mathbf{K}_1(H) \subseteq f^{-1}(\mathbf{K}_2(f(H)))$.

We now show that (iii) implies (iv). Let $V$ be a neighborhood of $f(x)$ in $(T, \mathcal{O}')$ and let $W$ be an open set such that $f(x) \in W \subseteq V$. Define the set $U \subseteq S$ as $U = S - f^{-1}(T - W)$. Since $f(x) \in W$, $f(x) \notin T - W$, $x \notin f^{-1}(T - W)$ and therefore $x \in U$.

By (iii), we have

$$f\left(\mathbf{K}_1(f^{-1}(T - W))\right) \subseteq \mathbf{K}_2(f(f^{-}1(T - W))) \subseteq \mathbf{K}_2(T - W) = T - W,$$

because $T - W$ is a closed set. Consequently, $\mathbf{K}_1(f^{-1}(T - W)) \subseteq f^{-1}(T - W)$, so $\mathbf{K}_1(f^{-1}(T - W)) = f^{-1}(T - W)$, which implies that $f^{-1}(T - W)$ is a closed set. This means that $U$ is an open set, and hence it is a neighborhood of $x$. Then,

$$f(U) = f\left(S - f^{-1}(T - W)\right) = f(f^{-1}(W)) \subseteq W.$$

Finally, to show that (iv) implies (i), let $V$ be an open set in $(T, \mathcal{O}')$ and $x \in f^{-1}(V)$, so $f(x) \in V$. Since $V$ is open, it is a neighborhood of $f(x)$, so by (iv) there exists $U \in neigh_x(\mathcal{O})$ such that $f(U) \subseteq V$, which implies $U \subseteq f^{-1}(V)$ and $f^{-1}(V)$ is a neighborhood of $x$. By Theorem 6.23, $f^{-1}(V)$ is open so $f$ is continuous. □

**Definition 6.65.** *Let $(S, \mathcal{O})$ and $(T, \mathcal{O}')$ be two topological spaces. A bijection $f : S \longrightarrow T$ is a* homeomorphism *if both $f$ and its inverse $f^{-1}$ are continuous functions.*

*If a homeomorphism exists between the topological spaces* $(S, \mathcal{O})$ *and* $(S, \mathcal{O}')$*, we say that these spaces are* homeomorphic.

**Theorem 6.66.** *A bijection* $f : S \longrightarrow T$ *between two topological spaces* $(S, \mathcal{O})$ *and* $(T, \mathcal{O}')$ *is a homeomorphism if and only if* $U \in \mathcal{O}$ *is equivalent to* $f(U) \in \mathcal{O}'$*.*

*Proof.* Suppose that $f$ is a homeomorphism between $(S, \mathcal{O})$ and $(T, \mathcal{O}')$. If $U \in \mathcal{O}$ the continuity of $f^{-1}$ implies that $(f^{-1})^{-1}(U) = f(U) \in \mathcal{O}'$; on the other hand, if $f(U) \in \mathcal{O}'$, then, since $U = f^{-1}(f(U))$, the continuity of $f$ yields $U \in \mathcal{O}$.

Conversely, suppose that for the bijection $f : S \longrightarrow T$, $U \in \mathcal{O}$ if and only if $f(U) \in \mathcal{O}'$. Suppose that $V \in \mathcal{O}'$; since $f$ is a bijection, there is $W \subseteq S$ such that $V = f(W)$ and $W \in \mathcal{O}$ by hypothesis. Observe that $f^{-1}(V) = W$, so $f$ is continuous. To prove that $f^{-1}$ is continuous, note that we need to verify that $(f^{-1})^{-1}(Z)$ is an open set in $(S, \mathcal{O})$ for any set $Z \in \mathcal{O}'$, which is effectively the case because $(f^{-1})^{-1}(Z) = f(Z)$. □

Any property of $(S, \mathcal{O})$ that can be expressed using the open sets of this topological space is preserved in topological spaces $(T, \mathcal{O}')$ that are homeomorphic to $(S, \mathcal{O})$. Therefore, such a property is said to be *topological*.

The collection of all pairs of topological spaces that are homeomorphic is an equivalence relation on the class of topological spaces as can be easily shown.

*Example 6.67.* We prove that all open intervals of $\mathbb{R}$, bounded or not, are homeomorphic.

Let $(a, b)$ and $(c, d)$ be two bounded intervals of $\mathbb{R}$ and let $f : (a, b) \longrightarrow (c, d)$ be the linear function defined by $f(x) = px + q$, where $p = \frac{d-c}{b-a}$ and $q = \frac{bc-ad}{b-a}$. It is easy to verify that $f$ is a homeomorphism, so any two bounded intervals of $\mathbb{R}$ are homeomorphic; in particular, any bounded interval $(a, b)$ is homeomorphic with $(0, 1)$.

Any two unbounded intervals $(a, \infty)$ and $(b, \infty)$ are homeomorphic; the mapping $g(x) = \frac{b}{a}x$ is a homeomorphism between these sets. Similarly, any two unbounded intervals of the form $(-\infty, a)$ and $(-\infty, b)$ are homeomorphic, and so are $(a, \infty)$ and $(-\infty, b)$.

The function $h : (0, 1) \longrightarrow (0, \infty)$ defined by $h(x) = \tan \frac{\pi x}{2}$ is a homeomorphism, whose inverse mapping is $h^{-1}(x) = \frac{2}{\pi} \arctan x$ so $(0, 1)$ is homeomorphic with $(0, \infty)$. Finally, $(-1, 1)$ is homeomorphic to $(-\infty, \infty)$ since the mapping $h_1 : (-1, 1) \longrightarrow (-\infty, \infty)$ defined by $h(x) = \tan \frac{\pi x}{2}$ for $x \in (-1, 1)$ is a homeomorphism.

The next theorem shows that compactness is preserved by continuous functions.

**Theorem 6.68.** *Let* $(S, \mathcal{O})$ *and* $(T, \mathcal{O}')$ *be two topological spaces and let* $f : S \longrightarrow T$ *be a continuous function. If* $(S, \mathcal{O})$ *is compact, then* $f(S)$ *is compact in* $(T, \mathcal{O}')$*.*

*Proof.* Let $\mathcal{D} = \{D_i \mid i \in I\}$ be an open cover of $f(S)$. Then $f^{-1}(D_i)$ is an open set in $(S, \mathcal{O})$ because $f$ is continuous and the collection $\mathcal{C} = \{f^{-1}(D_i) \mid i \in I\}$ is an open cover of $S$. Since $(S, \mathcal{O})$ is compact, there exists a finite subcover $\mathcal{C}_1 = \{f^{-1}(D_i) \mid i \in I_1\}$ of $S$ ($I_1$ is a finite subset of $I$). Since $S = \bigcup\{f^{-1}(D_i) \mid i \in I_1\}$, we have

$$\begin{aligned} f(S) &= f\left(\bigcup\{f^{-1}(D_i) \mid i \in I_1\}\right) \\ &= \bigcup\{f(f^{-1}(D_i)) \mid i \in I_1\} \\ &= \bigcup\{D_i \mid i \in I_1\}, \end{aligned}$$

which shows that $\mathcal{D}$ contains a finite subcover of $f(S)$. □

Using the notion of the neighborhood of an element, it is possible to localize the notion of continuity.

**Definition 6.69.** *Let $(S, \mathcal{O})$ and $(T, \mathcal{O}')$ be two topological spaces. A function $f : S \longrightarrow T$ is* continuous at *$s$, where $s \in S$, if for every neighborhood $V$ of $f(s)$ there exists a neighborhood $U$ of $s$ such that $f(U) \subseteq V$.*

**Theorem 6.70.** *Let $(S, \mathcal{O})$ and $(T, \mathcal{O}')$ be two topological spaces. A function $f : S \longrightarrow T$ is continuous if and only if it is continuous at every element $s$ of $S$.*

*Proof.* This statement follows immediately from Definition 6.70 and from the last part of Theorem 6.64. □

## 6.7 Connected Topological Spaces

We now discuss a formalization of the notion of a "one-piece" topological space.

**Theorem 6.71.** *Let $(S, \mathcal{O})$ be a topological space. The following statements are equivalent:*

(i) *There exists a clopen subset $K$ of $S$ such that $K \notin \{\emptyset, S\}$.*
(ii) *There exist two non-empty open subsets $L, L'$ of $S$ that are complementary.*
(iii) *There exist two non-empty closed subsets $H, H'$ of $S$ that are complementary.*

*Proof.* (i) implies (ii): If $K$ is clopen and $K \notin \{\emptyset, S\}$, then both $K$ and $\bar{K}$ are non-empty open sets.

(ii) implies (iii): Suppose that $L$ and $L'$ are two non-empty complementary open subsets of $S$. Then, $L$ and $L'$ are in the same time closed because the complements of each set is open.

(iii) implies (i): If $H$ and $H'$ are complementary closed sets, then each of them is also open because the complements of each set is closed. Thus, both sets are clopen. □

**Definition 6.72.** *A topological space $(S, \mathcal{O})$ is* disconnected *if it satisfies any of the equivalent conditions of Theorem 6.71. Otherwise, $(S, \mathcal{O})$ is said to be* connected.

*A subset $T$ of a connected topological space is* connected *if the subspace $T$ is connected.*

**Theorem 6.73.** *Let $T$ be a subset of $S$, where $(S, \mathcal{O})$ is a topological space. The following statements are equivalent:*

(i) *$T$ is connected.*

(ii) *There are no open sets $L_1, L_2$ in $(S, \mathcal{O})$ such that $T \subseteq L_1 \cup L_2$, and $T \cap L_1$, and $T \cap L_2$ are non-empty and disjoint.*

(iii) *There are no closed sets $H_1, H_2$ in $(S, \mathcal{O})$ such that $T \subseteq H_1 \cup H_2$, and $T \cap H_1$ and $T \cap H_2$ are non-empty and disjoint.*

(iv) *There is no clopen set in $(S, \mathcal{O})$ that has a non-empty intersection with $T$.*

*Proof.* The equivalence of the statements follows immediately from the definition of the subspace topology. □

**Theorem 6.74.** *Let $\mathcal{C} = \{C_i \mid i \in I\}$ be a family of connected subsets of a topological space $(S, \mathcal{O})$. If $C_i \cap C_j \neq \emptyset$ for every $i, j \in I$ such that $i \neq j$, then $\bigcup \mathcal{C}$ is connected.*

*Proof.* Suppose that $C = \bigcup \mathcal{C}$ is not connected. Then $C$ contains two complementary open subsets $L'$ and $L''$. For every $i \in I$, the sets $C_i \cap L'$ and $C_i \cap L''$ are complementary and open in $C_i$. Since each $C_i$ is connected, we have either $C_i \cap L' = \emptyset$ or $C_i \cap L'' = \emptyset$ for every $i \in I$. In the first case, $C_i \subseteq L''$, while in the second, $C_i \subseteq L'$. Thus, the collection $\mathcal{C}$ can be partitioned into two subcollections, $\mathcal{C} = \mathcal{C}' \cup \mathcal{C}''$, where $\mathcal{C}' = \{C_i \in \mathcal{C} \mid C_i \subseteq L'\}$ and $\mathcal{C}'' = \{C_i \in \mathcal{C} \mid C_i \subseteq L''\}$. Clearly, two sets $C_i \in \mathcal{C}'$ and $C_j \in \mathcal{C}''$ are disjoint because the sets $L'$ and $L''$ are disjoint, and this contradicts the hypothesis. □

**Corollary 6.75.** *Let $(S, \mathcal{O})$ be a topological space and let $x \in S$. The collection $\mathcal{C}_x$ of connected subsets of $S$ that contain $x$ has $K_x = \bigcup \mathcal{C}_x$ as its largest element.*

*Proof.* This follows immediately from Theorem 6.74. □

We will refer to $K_x$ as *the connected component* of $x$.

**Theorem 6.76.** *Let $T$ be a connected subset of a topological space $(S, \mathcal{O})$, and suppose that $W$ is a subset of $S$ such that $T \subseteq W \subseteq \boldsymbol{K}(T)$. Then $W$ is connected.*

*Proof.* Suppose that $W$ is not connected (that is, $W = U \cup U'$, where $U$ and $U'$ are two nonempty, disjoint, and open sets in $W$). There exist two open sets $L, L'$ in $S$ such that $U = W \cap L$ and $U' = W \cap L'$. Since $T \subseteq W$, the sets

$T \cap U$ and $T \cap U'$ are open in $T$, disjoint, and their union equals $T$. Thus, we have either $T \cap U = \emptyset$ or $T \cap U' = \emptyset$ because $T$ is connected.

If $T \cap U = \emptyset$, then $T \cap L = (T \cap W) \cap L = T \cap (W \cap L) = T \cap U = \emptyset$, so $T \subseteq \bar{L}$. Since $\bar{L}$ is closed, $\mathbf{K}(T) \subseteq \bar{L}$, which implies $W \subseteq \bar{L}$, which implies $U = W \cap L = \emptyset$. This contradicts the assumption made earlier about $U$. A similar contradiction follows from $T \cap U' = \emptyset$. Thus, $W$ is connected. □

**Corollary 6.77.** *If $T$ is a connected subset of a topological space $(S, \mathcal{O})$, then $\boldsymbol{K}(T)$ is also connected.*

*Proof.* This statement is a special case of Theorem 6.75. □

**Theorem 6.78.** *Let $(S, \mathcal{O})$ be a topological space. The collection of all connected components of $S$ is a partition of $S$ that consists of closed sets.*

*Proof.* Corollary 6.77 implies that each connected component $K_x$ is closed. Suppose that $K_x$ and $K_y$ are two connected components that are not disjoint. Then, by Theorem 6.74, $K_x \cup K_y$ is connected. Since $x \in K_x \cup K_y$, it follows that $K_x \cup K_y \subseteq K_x$ because $K_x$ is the maximal connected set that contains $x$, so $K_y \subseteq K_x$. Similarly, $K_x \subseteq K_y$, so $K_x = K_y$. □

*Example 6.79.* The topological space $(\mathbb{R}, \mathcal{O})$ is connected. Suppose that $K$ is a clopen set in $\mathbb{R}$ distinct from $\mathbb{R}$ and $\emptyset$, and let $x \in \mathbb{R} - K$.

Suppose that the set $K \cap [x, \infty)$ is nonempty. Then, this set is closed and bounded below and therefore has a least element $u$. Since $K \cap [x, \infty) = K \cap (x, \infty)$ is also open, there exists $\epsilon > 0$ such that $(u - \epsilon, u + \epsilon) \subseteq K \cap [x, \infty)$, which contradicts the fact that $u$ is the least element of $K \cap [x, \infty)$. A similar contradiction is obtained if we assume that $K \cap (-\infty, x] \neq \emptyset$, so $\mathbb{R}$ cannot contain a clopen set distinct from $\mathbb{R}$ or $\emptyset$.

**Theorem 6.80.** *The image of a connected topological space through a continuous function is a connected set.*

*Proof.* Let $(S_1, \mathcal{O}_1)$ and $(S_2, \mathcal{O}_2)$ be two topological spaces and let $f : S_1 \longrightarrow S_2$ be a continuous function, where $S_1$ is connected. If $f(S_1)$ were not connected, we would have two nonempty open subsets $L$ and $L'$ of $f(S_1)$ that are complementary. Then, $f^{-1}(L)$ and $f^{-1}(L')$ would be two nonempty, open sets in $S_1$ which are complementary, which contradicts the fact that $S_1$ is connected. □

A characterization of connected spaces is given next.

**Theorem 6.81.** *Let $(S, \mathcal{O})$ be a topological space and let $(\{0, 1\}, \mathcal{P}(\{0, 1\})$ be a two-element topological space equipped with the non-discrete topology. Then, $S$ is connected if and only if every continuous application $f : S \longrightarrow \{0, 1\}$ is constant.*

*Proof.* Suppose that $S$ is connected. Both $f^{-1}(0)$ and $f^{-1}(1)$ are clopen sets in $S$ because both $\{0\}$ and $\{1\}$ are clopen in the discrete topology. Thus, we have either $f^{-1}(0) = \emptyset$ and $f^{-1}(1) = S$, or $f^{-1}(0) = S$ and $f^{-1}(1) = \emptyset$. In the first case, $f$ is the constant function $f(x) = 1$; in the second, it is the constant function $f(x) = 0$.

Conversely, suppose that the condition is satisfied for every continuous function $f : S \longrightarrow \{0, 1\}$ and suppose $(S, \mathcal{O})$ is not connected. Then, there exist two nonempty disjoint open subsets $L$ and $L'$ that are complementary. Let $f = 1_L$ be the indicator function of $L$, which is continuous because both $L$ and $L'$ are open. Thus, $f$ is constant and this implies either $L = \emptyset$ and $L' = S$ or $L = S$ and $L' = \emptyset$, so $S$ is connected. □

*Example 6.82.* Theorem 6.81 allows us to prove that the connected subsets of $\mathbb{R}$ are exactly the intervals.

Suppose that $T$ is a connected subset of $S$ but is not an interval. Then, there are three numbers $x, y, z$ such that $x < y < z$, $x, z \in T$ but $y \notin T$. Define the function $f : T \longrightarrow \{0, 1\}$ by $f(u) = 0$ if $u < y$ and $f(u) = 1$ if $y < u$. Clearly, $f$ is continuous but is not constant, and this contradicts Theorem 6.81. Thus, $T$ must be an interval.

Suppose now that $T$ is an open interval of $\mathbb{R}$. We saw that $T$ is homeomorphic to $\mathbb{R}$ (see Example 6.67), so $T$ is indeed connected. If $T$ is an arbitrary interval, its interior $\mathbf{I}(T)$ is an open interval and, since $\mathbf{I}(T) \subseteq T \subseteq \mathbf{K}(\mathbf{I}(T))$, it follows that $T$ is connected.

**Definition 6.83.** *A topological space $(S, \mathcal{O})$ is* totally disconnected *if, for every $x \in S$, the connected component of $x$ is $K_x = \{x\}$.*

*Example 6.84.* Any topological space equipped with the discrete topology is totally disconnected.

**Theorem 6.85.** *Let $(S, \mathcal{O})$ be a topological space and let $T$ be a subset of $S$.*

*If for every pair of distinct points $x, y \in T$ there exist two disjoint closed sets $H_x$ and $H_y$ such that $T \subseteq H_x \cup H_y$, $x \in H_x$, and $y \in H_y$, then $T$ is totally disconnected.*

*Proof.* Let $K_x$ be the connected component of $x$, and suppose that $y \in K_x$ and $y \neq x$, that is, $K_x = K_y = K$. Then, $K \cap H_x$ and $K \cap H_y$ are nonempty disjoint closed sets and $K = (K \cap H_x) \cup (K \cap H_y)$, which contradicts the connectedness of $K$. Therefore, $K_x = \{x\}$ for every $x \in T$ and $T$ is totally disconnected. □

## 6.8 Separation Hierarchy of Topological Spaces

The next definition introduces a hierarchy of topological spaces that is based on *separation properties* of these spaces.

**Definition 6.86.** *Let $(S, \mathcal{O})$ be a topological space and let $x$ and $y$ be two arbitrary, distinct elements of $S$. This topological space is:*

(i) a $T_0$ space *if there exists $U \in \mathcal{O}$ such that one member of the set $\{x, y\}$ belongs to $U$ and the other to $S - U$;*

(ii) a $T_1$ space *if there exist $U, V \in \mathcal{O}$ such that $x \in U - V$ and $y \in V - U$;*

(iii) a $T_2$ space *or a* Hausdorff space *if there exist $U, V \in \mathcal{O}$ such that $x \in U$ and $y \in V$ and $U \cap V = \emptyset$;*

(iv) a $T_3$ space *if for every closed set $H$ and $x \in S - H$ there exist $U, V \in \mathcal{O}$ such that $x \in U$ and $H \subseteq V$ and $U \cap V = \emptyset$;*

(v) a $T_4$ space *if for all disjoint closed sets $H, L$ there exist $U, V \in \mathcal{O}$ such that $H \subseteq U$, $L \subseteq V$, and $U \cap V = \emptyset$.*

**Theorem 6.87.** *A topological space $(S, \mathcal{O})$ is a $T_1$ space if and only if every singleton $\{x\}$ is a closed set.*

*Proof.* Suppose that $(S, \mathcal{O})$ is a $T_1$, space and for every $y \in S - \{x\}$ let $U_y$ and $V_y$ be two open sets such as $x \in U_y - V_y$ and $y \in V_y - U_y$. Then, $x \in \bigcup_{y \neq x} U_y$ and $x \notin \bigcup_{y \neq x} V_y$, so $y \in \bigcup_{y \neq x} V_y \subseteq S - \{x\}$. Thus, $S - \{x\}$ is an open set, so $\{x\}$ is closed.

Conversely, suppose that each singleton $\{u\}$ is closed. Let $x, y \in S$ be two distinct elements of $S$. Note that the sets $S - \{x\}$ and $S - \{y\}$ are open and $x \in (S - \{y\}) - (S - \{x\})$ and $y \in (S - \{x\}) - (S - \{y\})$, which shows that $(S, \mathcal{O})$ is a $T_1$-space. □

**Theorem 6.88.** *Let $(S, \mathcal{O})$ be a $T_4$-separated topological space. If $H$ is a closed set and $L$ is an open set such that $H \subseteq L$, then there exists an open set $U$ such that $H \subseteq U \subseteq \boldsymbol{K}(U) \subseteq L$.*

*Proof.* Observe that $H$ and $S - L$ are two disjoint closed sets under the assumptions of the theorem. Since $(S, \mathcal{O})$ is a $T_4$-separated topological space, there exist $U, V \in \mathcal{O}$ such that $H \subseteq U$, $S - L \subseteq V$ and $U \cap V = \emptyset$. This implies $U \subseteq S - V \subseteq L$. Since $S - V$ is closed, we have

$$H \subseteq U \subseteq \mathbf{K}(U) \subseteq \mathbf{K}(S - V) = S - V \subseteq L,$$

which proves that $U$ satisfies the conditions of the theorem. □

The next theorem is in some sense a reciprocal result of Theorem 6.60, which holds in the realm of Hausdorff spaces.

**Theorem 6.89.** *Each compact subset of a Hausdorff space $(S, \mathcal{O})$ is closed.*

*Proof.* Let $H$ be a compact subset of $(S, \mathcal{O})$ and let $y$ be an element of the set $S - H$. It suffices to show that the set $S - H$ is open. For every $x \in H$, we have two open subsets $U_x$ and $V_x$ such that $x \in U_x$, $y \in V_x$ and $U_x \cap V_x = \emptyset$. The collection $\{U_x \mid x \in H\}$ is an open cover of $H$ and the compactness of $H$ implies the existence of a finite subcover $U_{x_1}, \ldots, U_{x_n}$ of $H$. Consider the

open set $V = \bigcap_{i=1}^{n} V_{x_i}$, which is disjoint from each of the sets $U_{x_1}, \ldots, U_{x_n}$ and, therefore, it is disjoint from $H$. Thus, for every $y \in S - H$ there exists an open set $V$ such that $y \in V \subseteq S - H$, which implies that $S - H$ is open. □

**Corollary 6.90.** *In a Hausdorff space $(S, \mathcal{O})$, each finite subset is a closed set.*

*Proof.* Since every finite subset of $S$ is compact, the statement follows immediately from Theorem 6.89. □

It is clear that every $T_2$ space is a $T_1$ space and each $T_1$ space is a $T_0$ space. However, this hierarchy does not hold beyond $T_2$. This requires the introduction of two further classes of topological spaces.

**Definition 6.91.** *A topological space $(S, \mathcal{O})$ is* regular *if it is both a $T_1$ and a $T_3$ space; $(S, \mathcal{O})$ is* normal *if it is both a $T_1$ and a $T_4$ space.*

**Theorem 6.92.** *Every regular topological space is a $T_2$ space and every normal topological space is a regular one.*

*Proof.* Let $(S, \mathcal{O})$ be a topological space that is regular and let $x$ and $y$ be two distinct points in $S$. By Theorem 6.87, the singleton $\{y\}$ is a closed set. Since $(S, \mathcal{O})$ is a $T_3$, space, two open sets $U$ and $V$ exist such that $x \in U$, $\{y\} \subseteq V$, and $U \cap V = \emptyset$, so $(S, \mathcal{O})$ is a $T_2$ space. We leave the second part of the theorem to the reader. □

## 6.9 Products of Topological Spaces

**Theorem 6.93.** *Let $\{(S_i, \mathcal{O}_i) \mid i \in I\}$ be a family of topological spaces indexed by the set $I$. Define on the set $S = \prod_{i \in I} S_i$ the collection of sets $\mathcal{B} = \{\bigcap_{j \in J} p_j^{-1}(L_j) \mid L_j \in \mathcal{O}_j$ and $J$ finite$\}$. Then, $\mathcal{B}$ is a basis.*

*Proof.* Note that every set $\bigcap_{j \in J} p_j^{-1}(L_j)$ has the form $\prod_{i \in I - J} \times \prod_{j \in J} L_j$. We need to observe only that a finite intersection of sets in $\mathcal{B}$ is again a set in $\mathcal{B}$. Therefore, $\mathcal{B}$ is a basis. □

**Definition 6.94.** *The topology $TOP(\mathcal{B})$ generated on the set $S$ by $\mathcal{B}$ is called the* product of the topologies $\mathcal{O}_i$ *and is denoted by $\prod_{i \in I} \mathcal{O}_i$.*

*The topological space $\{(S_i, \mathcal{O}_i) \mid i \in I\}$ is the* product of the collection of topological spaces $\{(S_i, \mathcal{O}_i) \mid i \in I\}$.

The product of the topologies $\{\mathcal{O}_i \mid i \in I\}$ can be generated starting from the subbasis $\mathcal{S}$ that consists of sets of the form $D_{j,L} = \{t \mid t \in \prod_{i \in I} \mid t(j) \in L\}$, where $j \in I$ and $L$ is an open set in $(S_j, \mathcal{O}_j)$. It is easy to see that any set in the basis $\mathcal{B}$ is a finite intersection of sets of the form $D_{j,L}$.

*Example 6.95.* Let $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$, where the product involves $n$ copies of $\mathbb{R}$ and $n \geq 1$. In Example 6.44, we saw that the collection of open intervals $\{(a, b) \mid a, b \in \mathbb{R} \text{ and } a < b\}$ is a basis for the topological space $(\mathbb{R}, \mathcal{O})$. Therefore, a basis of the topological space $(\mathbb{R}^n, \mathcal{O} \times \cdots \mathcal{O})$ consists of parallelepipeds of the form $(a_1, b_1) \times \cdots \times (a_n, b_n)$, where $a_i < b_i$ for $1 \leq i \leq n$.

**Theorem 6.96.** *Let $\{(S_i, \mathcal{O}_i) \mid i \in I\}$ be a collection of topological spaces. Each projection $p_\ell : \prod_{i \in I} S_i \longrightarrow S_\ell$ is a continuous function for $\ell \in I$. Moreover, the product topology is the coarsest topology on $S$ such that projections are continuous.*

*Proof.* Let $L$ be an open set in $(S_\ell, \mathcal{O}_\ell)$. We have

$$p_\ell^{-1}(L) = \left\{ t \in \prod_{i \in I} S_i \mid t(\ell) \in L \right\},$$

which has the form $\prod_{i \in I} K_i$, where each set $K_i$ is open because

$$K_i = \begin{cases} S_i & \text{if } i \neq \ell, \\ L & \text{if } i = \ell, \end{cases}$$

for $i \in I$. Thus, $p_\ell^{-1}(L)$ is open and $p_\ell$ is continuous.

The proof of the second part of the theorem is left to the reader. □

The next lemma is a preliminary result to a theorem that refers to the compactness of products of topological spaces.

**Lemma 6.97.** *Let $\mathcal{C}$ be a collection of subsets of $S = \prod_{i \in I} S_i$ such that $\mathcal{C}$ has the f.i.p. and $\mathcal{C}$ is maximal with this property.*

*We have $\bigcap \mathcal{D} \in \mathcal{C}$ for every finite subcollection $\mathcal{D}$ of $\mathcal{C}$. Furthermore, if $T \cap C \neq \emptyset$ for every $C \in \mathcal{C}$, then $T \in \mathcal{C}$.*

*Proof.* Let $\mathcal{D} = \{D_1, \ldots, D_n\}$ be a finite subcollection of $\mathcal{C}$ and let $D = \bigcap \mathcal{D} \neq \emptyset$. Note that the intersection of every finite subcollection of $\mathcal{C} \cup \{D\}$ is also nonempty. The maximality of $\mathcal{C}$ implies $D \in \mathcal{C}$, which proves the first part of the lemma.

For the second part of the lemma, observe that the intersection of any finite subcollection of $\mathcal{D} \cup \{T\}$ is not empty. Therefore, as above, $T \in \mathcal{C}$. □

**Theorem 6.98 (Tychonoff's Theorem).** *Let $\{(S_i, \mathcal{O}_i) \mid i \in I\}$ be a collection of topological spaces such that $S_i \neq \emptyset$ for every $i \in I$. Then, $(\prod_{i \in I} S_i, \prod_{i \in I} \mathcal{O}_i)$ is compact if and only if each topological space $(S_i, \mathcal{O}_i)$ is compact for $i \in I$.*

*Proof.* If $(\prod_{i \in I} S_i, \mathcal{O})$ is compact, then, by Theorem 6.68, it is clear that each of the topological spaces $(S_i, \mathcal{O}_i)$ is compact because each projection $p_i$ is continuous.

Conversely, suppose that each of the topological spaces $(S_i, \mathcal{O}_i)$ is compact.

Let $\mathcal{E}$ be a family of sets in $S = \prod_{i \in I} S_i$ that has the f.i.p. and let $(\mathfrak{C}, \subseteq)$ be the partially ordered set whose elements are collections of subsets of $S$ that have the f.i.p. and contain the family $\mathcal{E}$.

Let $\{\mathcal{C}_i \mid i \in I\}$ be a chain in $(\mathfrak{C}, \subseteq)$. It is easy to verify that $\bigcup\{\mathcal{C}_i \mid i \in I\}$ has the f.i.p., so every chain in $(\mathfrak{C}, \subseteq)$ has an upper bound. Therefore, by Zorn's lemma (see Theorem 4.101), the poset $(\mathfrak{C}, \subseteq)$ contains a maximal collection $\mathcal{C}$ that has the f.i.p.and contains $\mathcal{E}$. We aim to find an element $t \in \prod_{i \in I} S_i$ that belongs to $\bigcap\{\mathbf{K}(C) \mid C \in \mathcal{C}\}$ because, in this case, the same element will belong to $\bigcap\{\mathbf{K}(C) \mid C \in \mathcal{E}\}$ and this would imply by Theorem 6.53 that $(S, \mathcal{O})$ is compact.

Let $\mathcal{C}_i$ be the collection of closed subsets of $S_i$ defined by

$$\mathcal{C}_i = \{\mathbf{K}_i(p_i(C)) \mid C \in \mathcal{C}\}$$

for $i \in I$, where $\mathbf{K}_i$ is the closure of the topological space $(S_i, \mathcal{O}_i)$.

It is clear that each collection $\mathcal{C}_i$ has the f.i.p. in $S_i$. Indeed, since $\mathcal{C}$ has the f.i.p., if $\{C_1, \ldots, C_n\} \subseteq \mathcal{C}$ and $x \in \bigcap_{k=1}^n C_k$, then $p_i(x) \in \bigcap_{k=1}^n \mathbf{K}(p_i(C_k))$, so $\mathcal{C}_i$ has the f.i.p. Since $(S_i, \mathcal{O}_i)$ is compact, we have $\bigcap \mathcal{C}_i \neq \emptyset$, by Part (iii) of Theorem 6.52. Let $t_i \in \bigcap \mathcal{C}_i = \bigcap\{\mathbf{K}_i(p_i(C)) \mid C \in \mathcal{C}\}$ and let $t \in S$ be defined by $t(i) = t_i$ for $i \in I$.

Let $D_{j,L} = \{u \mid u \in \prod_{i \in I} \mid u(j) \in L\}$, a set of the subbasis of the product topology that contains $t$, defined earlier, where $L$ is an open set in $(S_j, \mathcal{O}_j)$. Since $g(j) \in L$, the set $L$ has a nonempty intersection with every set $\mathbf{K}_i(p_i(C))$, where $C \in \mathcal{C}$. On the other hand, since $p_i(D_{j,L}) = S_i$ for $i \neq j$, it follows that for every $i \in I$ we have $p_i(D_{j,L}) \cap \bigcap_{C \in \mathcal{C}} \mathbf{K}_i(p_i(C)) \neq \emptyset$. Therefore, $p_i(D_{j,L})$ has a nonempty intersection with every set of the form $\mathbf{K}_i(p_i(C))$, where $C \in \mathcal{C}$. By the contrapositive of Theorem 6.8, this means that $p_i(D_{j,L}) \cup p_i(C) \neq \emptyset$ for every $i \in I$ and $C \in \mathcal{C}$. This in turn means that $D_{j,L} \cup C \neq \emptyset$ for every $C \in \mathcal{C}$. By Lemma 6.97, it follows that $D_{j,L} \in \mathcal{C}$. Since every set that belongs to the basis of the product topology is a finite intersection of sets of the form $D_{j,L}$, it follows that any member of the basis has a nonempty intersection with every set of $\mathcal{C}$. This implies that $g$ belongs to $\bigcup\{\mathbf{K}(C) \mid C \in \mathcal{C}\}$, which implies the compactness of $(\prod_{i \in I} S_i, \prod_{i \in I} \mathcal{O}_i)$. □

*Example 6.99.* In Example 6.56, we have shown that every closed interval $[x, y]$ of $\mathbb{R}$ where $x < y$ is compact. By Theorem 6.98, any subset of $\mathbb{R}^n$ of the form $[x_1, y_1] \times \cdots \times [x_n, y_n]$ is compact.

## 6.10 Fields of Sets

In this section, we introduce collections of sets that play an important role in measure and probability theory.

**Definition 6.100.** *Let $S$ be a set. A* field of sets on $S$ *is a family of subsets $\mathcal{E}$ of $S$ that satisfies the following conditions:*
(i) $S \in \mathcal{E}$.
(ii) *If $U \in \mathcal{E}$, then $\bar{U} = S - U \in \mathcal{E}$.*
(iii) *If $U_0, \ldots, U_{n-1}$ belong to $\mathcal{E}$, then $\bigcup_{0 \leq i \leq n-1} U_i$ belongs to $\mathcal{E}$.*

*A* $\sigma$-field of sets *on $S$ is a family of subsets $\mathcal{E}$ of $S$ that satisfies Conditions (i) and (ii) and, in addition, satisfies the following condition:*
*(iii′) if $\{U_i \mid i \in \mathbb{N}\}$ is a countable family of sets included in $\mathcal{E}$, then $\bigcup_{i \in \mathbb{N}} U_i$ belongs to $\mathcal{E}$.*
*Clearly, every $\sigma$-field is also a field on $S$.*

*If $\mathcal{E}$ is a $\sigma$-field of sets on $S$, we shall refer to the pair $(S, \mathcal{E})$ as a* measurable space.

*Example 6.101.* The collection $\mathcal{E}_0 = \{\emptyset, S\}$ is a $\sigma$-field on $S$; moreover, for every $\sigma$-field $\mathcal{E}$ on $S$, we have $\mathcal{E}_0 \subseteq \mathcal{E}$.

The set $\mathcal{P}(S)$ of all subsets of a set $S$ is a $\sigma$-field on $S$.

If $T$ is a subset of $S$, then the collection $\{\emptyset, T, S - T, S\}$ is a $\sigma$-field on $S$.

**Theorem 6.102.** *The class of all fields ($\sigma$-fields) of sets on $S$ is a closure system on $\mathcal{P}(S)$.*

*Proof.* Let $\mathfrak{E} = \{\mathcal{E}_i \mid i \in I\}$ be a collection of fields of sets on $S$. Since $S \in \mathcal{E}_i$ for every $i \in I$, it follows that $S \in \bigcap\{\mathcal{E}_i \mid i \in I\}$.

Suppose that $A \in \bigcap \mathfrak{E}$. Since $A \in \mathcal{E}_i$ for every $i \in I$, it follows that $\bar{A} \in \mathcal{E}_i$ for every $i \in I$, which implies that $\bar{A} \in \bigcap\{\mathcal{E}_i \mid i \in I\}$.

Finally, if $\{A_i \mid 1 \leq i \leq n\} \in \bigcap\{\mathcal{E}_i \mid i \in I\}$, it is easy to see that $\bigcup_{i=1}^{n} A_i \in \bigcap\{\mathcal{E}_i \mid i \in I\}$.

A similar argument proves that the class of all $\sigma$-fields of sets is also a closure system on $\mathcal{P}(S)$. □

*Example 6.103.* Let $A$ be a subset of the set $S$. The $\sigma$-field generated by the collection $\{A\}$ is $\{\emptyset, A, \bar{A}, S\}$.

**Definition 6.104.** *Let $(S, \mathcal{O})$ be a topological space. A subset $T$ of $S$ is said to be a* Borel set *if it belongs to the $\sigma$-field generated by the topology $\mathcal{O}$.*

*The $\sigma$-field of Borel sets of $(S, \mathcal{O})$ is denoted by $\mathcal{B}_{\mathcal{O}}$.*

It is clear that all open sets are Borel sets. Also, every closed set, as a complement of an open set, is a Borel set.

*Example 6.105.* We identify several families of Borel subsets of the topological space $(\mathbb{R}, \mathcal{O})$.

It is clear that every open interval $(a, b)$ and every set $(a, \infty)$ or $(-\infty, a)$ is a Borel set for $a, b \in \mathbb{R}$ because they are open sets. The closed intervals of the form $[a, b]$ are Borel sets because they are closed sets in the topological space.

Since $[a, b) = (-\infty, b) - (-\infty, a)$, it follows that the half-open intervals of this form are also Borel sets.

For every $a \in \mathbb{R}$, we have $\{a\} \in \mathcal{B}_{\mathcal{O}}$ because $\{a\} = [a, b) - (a, b)$ for every $b \in \mathbb{R}$ such that $b > a$. Therefore, every countable subset $\{a_n \mid n \in \mathbb{N}\}$ of $\mathbb{R}$ is a Borel set.

*Example 6.106.* Let $\pi = \{B_i \mid i \in I\}$ be a countable partition of a set $S$. The $\sigma$-field generated by $\pi$ is

$$\mathcal{E}_\pi = \left\{ \bigcup_{i \in J} B_i \mid J \subseteq I \right\}.$$

Clearly, every block $B_i$ belongs to $\mathcal{E}_\pi$, so $\pi \subseteq \mathcal{E}_\pi$.

To verify that $\mathcal{E}_\pi$ is a $\sigma$-field, note first that we have $S \in \mathcal{E}_\pi$ since $S = \bigcup_{i \in I} B_i$. If $A \in \mathcal{E}_{pi}$, then $A = \bigcup_{i \in J} B_i$ for some subset $J$ of $I$, so $\bar{A} = \bigcup_{i \in I - J} B_i$, which shows that $\bar{A} \in \mathcal{E}_\pi$. Let $\{A_\ell \mid \ell \in L\}$ be a family of sets included in $\mathcal{E}_\pi$. For each set $A_\ell$, there exists a set $J_\ell$ such that $A_\ell = \bigcup \{B_i \mid i \in J_\ell\}$. Therefore,

$$\bigcup_{\ell \in L} A_\ell = \bigcup \left\{ B_i \mid i \in \bigcup_{\ell \in L} J_\ell \right\},$$

which shows that $\bigcup_{\ell \in L} A_\ell \in \mathcal{E}_\pi$. This proves that $\mathcal{E}_\pi$ is a $\sigma$-field. Moreover, any $\sigma$-field on $S$ that includes $\pi$ also includes $\mathcal{E}_\pi$, which concludes the argument.

**Theorem 6.107.** *Let $(S, \mathcal{E})$ be a measurable space. The following statements hold:*

(i) $\emptyset \in \mathcal{E}$.
(ii) *If* $\{A_i \mid i \in \mathbb{N}\} \subseteq \mathcal{E}$, *then* $\bigcap_{i \in \mathbb{N}} A_i \in \mathcal{E}$.
(iii) *If* $A, B \in \mathcal{E}$, *then* $A - B$ *and* $A \oplus B$ *belong to* $\mathcal{E}$.

*Proof.* The first statement follows from the fact that $\emptyset = \bar{S}$.

Let $\{A_i \mid i \in \mathbb{N}\}$ be a family of subsets of $S$ such that $A_i \in \mathcal{E}$ for $i \in \mathbb{N}$. Since $\overline{A_i} \in \mathcal{E}$, we have $\bigcup \{\overline{A_i} \mid i \in \mathbb{N}\} \in \mathcal{E}$. Thus,

$$\overline{\bigcup \{\overline{A_i} \mid i \in \mathbb{N}\}} = \bigcap \{A_i \mid i \in \mathbb{N}\} \in \mathcal{E},$$

which yields the second part of the theorem.

The third statement of the theorem is immediate. □

**Corollary 6.108.** *Let $(S, \mathcal{E})$ be a measurable space and let $\{U_n \mid n \in \mathcal{E}\}$ be a sequence of members of $\mathcal{E}$. Then, both* $\liminf\{U_n \mid n \in \mathbb{N}\}$ *and* $\limsup\{U_n \mid n \in \mathbb{N}\}$ *belong to* $\mathcal{E}$.

*Proof.* This statement follows immediately from Definition 6.100 and from Theorem 6.107. □

Note that if $(S, \mathcal{E})$ is a measurable space (that is, if $\mathcal{E}$ is a $\sigma$-field on $S$), then condition (iii′) of Definition 6.100 amounts to $\mathcal{E}_\sigma \subseteq \mathcal{E}$. Moreover, by Part (ii) of Theorem 6.107, we also have $\mathcal{E}_\delta \subseteq \mathcal{E}$.

*Example 6.109.* Let $S$ be an arbitrary set and let $\mathcal{B}$ be the family of sets that consists of sets that are either countable or complements of countable sets. We claim that $(S, \mathcal{B})$ is a measurable space.

Note that $S \in \mathcal{B}$ because $S$ is the complement of $\emptyset$, which is countable. Next, if $A \in \mathcal{B}$ is countable, $\bar{A}$ is a complement of a countable set, so $\bar{A} \in \mathcal{B}$; otherwise, if $A$ is not countable, then it is the complement of a countable set, which means that $\bar{A}$ is countable, so $\bar{A} \in \mathcal{B}$.

Let $A$ and $B$ be two sets of $\mathcal{B}$. If both are countable, then $A \cup B \in \mathcal{B}$. If $\bar{A}$ and $\bar{B}$ are countable, then $\overline{A \cup B} = \overline{A} \cap \overline{B}$, so $A \cup B \in \mathcal{B}$ because it has a countable complement. If $A$ is countable and $\overline{B}$ is countable, then $\overline{A} \cap \overline{B}$ is countable because it is a subset of $\overline{B}$. Therefore, $A \cup B \in \mathcal{B}$ as a complement of a countable set. The case where $\overline{A}$ and $B$ are countable is treated similarly. Thus, in any case, the union of two sets of $\mathcal{B}$ belongs to $\mathcal{B}$.

Finally, we have to prove that if $\{A_i \mid i \in \mathbb{N}\}$ is a family of sets included in $\mathcal{B}$, then the set $A = \bigcup_{i \in \mathbb{N}} A_i$ belongs to $\mathcal{B}$. Indeed, let us split the set $I$ into $I'$ and $I''$, where $i \in I'$ if the set $A_i$ is countable and $i \in I''$ if the complement $\overline{A_i} = S - A_i$ is countable. Note that both $A' = \bigcup_{i \in I'} A_i$ and $A'' = \bigcap_{i \in I''} \overline{A_i}$ are countable sets (by Theorem 1.130 and by the fact that every subset of a countable set is countable, respectively), and that $A = A' \cup \overline{A''}$. Since both $A'$ and $\overline{A''}$ belong to $\mathcal{B}$, it follows that $A \in \mathcal{B}$.

**Definition 6.110.** *Let $(S, \mathcal{D})$ and $(T, \mathcal{E})$ be two measurable spaces. A function $f : S \longrightarrow T$ is said to be* measurable *if $f^{-1}(V) \in \mathcal{D}$ for every $V \in \mathcal{E}$.*

It is easy to verify that if $(S_i, \mathcal{E}_i)$ are measurable spaces for $1 \leq i \leq 3$ and $f : S_1 \longrightarrow S_2$, $g : S_2 \longrightarrow S_3$ are measurable functions, then their composition $gf$ is also a measurable function.

**Theorem 6.111.** *Let $S$ and $T$ be two sets and let $f : S \longrightarrow T$ be a function. If $\mathcal{E}$ is a $\sigma$-field on $T$, then the collection $f^{-1}(\mathcal{E})$ defined by $f^{-1}(\mathcal{E}) = \{f^{-1}(V) \mid V \in \mathcal{E}\}$ is a $\sigma$-field on $S$.*

*Proof.* Since $T \in \mathcal{E}$, it is clear that $S = f^{-1}(T)$ belongs to $f^{-1}(\mathcal{E})$.

Suppose that $U \in f^{-1}(\mathcal{E})$; that is, $U = f^{-1}(W)$ for some $W \in \mathcal{E}$. Since $S - U = f^{-1}(T) - f^{-1}(W) = f^{-1}(T - W)$ (by Theorem 1.67), it follows that $S - U \in f^{-1}(\mathcal{E})$. Similarly, if $\{W_i \mid i \in \mathbb{N}\}$ is a countable family of sets included in $\mathcal{E}$, then $\{f^{-1}(W_i) \mid i \in \mathbb{N}\}$ is a countable family of sets included in $f^{-1}(\mathcal{E})$ and $\bigcup\{f^{-1}(W_i) \mid i \in \mathbb{N}\}$ belongs to $f^{-1}(\mathcal{E})$ by Theorem 1.65. Thus, $f^{-1}(\mathcal{E})$ is a $\sigma$-field on $S$. □

**Corollary 6.112.** *Let $f : S \longrightarrow T$ be a function, where $(T, \mathcal{E})$ is a measurable space. Then, $f^{-1}(\mathcal{E})$ is the least $\sigma$-field of subsets of $S$ such that $f$ is is a measurable function between $S$ and $(T, \mathcal{E})$.*

*Proof.* Suppose that $\mathcal{D}$ is a $\sigma$-field on $S$ such that $f$ is measurable. Then, $f^{-1}(E) \in \mathcal{D}$ for every $E \in \mathcal{E}$, so $f^{-1}(\mathcal{E}) \subseteq \mathcal{D}$. The statement follows immediately since $f^{-1}(\mathcal{E})$ is a $\sigma$-field of sets. □

**Theorem 6.113.** *Let $S$ and $T$ be two sets and let $f : S \longrightarrow T$ be a function. If $\mathcal{E}$ is a $\sigma$-field on $S$, then the collection $\mathcal{E}' = \{W \in \mathcal{P}(T) \mid f^{-1}(W) \in \mathcal{E}\}$ is a $\sigma$-field on $T$.*

*Proof.* The proof is straightforward and is left to the reader as an exercise. □

**Theorem 6.114.** *Let $(S, \mathcal{O})$ and $(T, \mathcal{O}')$ be two topological spaces and let $f : S \longrightarrow T$ be a continuous function. Then, $f$ is measurable relative to the measurable spaces $(S, \mathcal{B}_{\mathcal{O}})$ and $(T, \mathcal{B}_{\mathcal{O}'})$, where $\mathcal{B}_{\mathcal{O}}$ and $\mathcal{B}_{\mathcal{O}'}$ are the collections of Borel sets in $(S, \mathcal{O})$ and $(T, \mathcal{O}')$, respectively.*

*Proof.* The collection of sets $\mathcal{E}' = \{W \in \mathcal{P}(T) \mid f^{-1}(W) \in \mathcal{B}_{\mathcal{O}}\}$ is a $\sigma$-field on $T$. Since $f$ is continuous, it is clear that $\mathcal{E}'$ contains every open set in $\mathcal{O}'$, so the $\sigma$-field of Borel sets $\mathcal{B}_{\mathcal{O}'}$ that is generated by $\mathcal{O}'$ is contained in $\mathcal{E}'$. Thus, for every Borel set $U$ in $T$, $f^{-1}(U) \in \mathcal{B}_{\mathcal{O}}$, which allows us to conclude that $f$ is indeed measurable. □

Next, we describe the $\sigma$-field generated by a countable partition of a set.

**Theorem 6.115.** *Let $\pi = \{B_i \mid i \in I\}$ be a countable partition of a set $S$. In other words, we assume that the set of indices $I$ of the blocks of $\pi$ is countable.*

*The $\sigma$-field generated by $\pi$ is the collection of sets:*

$$\left\{\bigcup_{i \in J} B_i \mid J \subseteq I\right\}.$$

*Proof.* Let $\mathcal{E}_\pi$ be the $\sigma$-field generated by $\pi$. Clearly, we have

$$\pi \subseteq \left\{\bigcup_{i \in J} B_i \mid J \subseteq I\right\} \subseteq \mathcal{E}_\pi.$$

The collection $\{\bigcup_{i \in J} B_i \mid J \subseteq I\}$ is a $\sigma$-field. Indeed, we have $S = \bigcup\{B \mid B \in \pi\}$, so $S \in \{\bigcup_{i \in J} B_i \mid J \subseteq I\}$.

Suppose that $A = \bigcup\{B_i \mid i \in J\}$. Then $\bar{A} = \{B_i \mid i \in I - J\}$, which shows that $\bar{A} \in \{\bigcup_{i \in J} B_i \mid J \subseteq I\}$.

Suppose that $A_0, \ldots, A_n, \ldots$ belong to $\mathcal{E}$, so $A_k = \bigcup\{B_i \mid i \in J_k\}$, where $J_k \subseteq I$ for $k \in \mathbb{N}$. Then, $\bigcup_{k \geq 0} A_k = \bigcup\{B_i \mid i \in \bigcup_{k \geq 0} J_k\}$, which implies that $\bigcup_{k \geq 0} A_k \in \{\bigcup_{i \in J} B_i \mid J \subseteq I\}$.

This implies that $\mathcal{E}_\pi = \{\bigcup_{i \in J} B_i \mid J \subseteq I\}$. □

We now give a technical result that concerns $\sigma$-fields.

**Theorem 6.116.** *Let $(S, \mathcal{E})$ be a measurable space and let $\{U_i \in \mathcal{E} \mid i \in \mathbb{N}\}$ be a family of sets from $\mathcal{E}$. There exists a family of sets $\{V_i \in \mathcal{E} \mid i \in \mathbb{N}\}$ that satisfies the following conditions:*
(i) *If $i, j \in \mathbb{N}$ and $i \neq j$, then $V_i \cap V_j = \emptyset$.*
(ii) *$V_i \subseteq U_i$ for $i \in \mathbb{N}$.*
(iii) $\bigcup\{V_i \mid i \in \mathbb{N}\} = \bigcup\{U_i \mid i \in \mathbb{N}\}$.

*Proof.* The sets $V_n$ are defined inductively by

$$V_0 = U_0,$$
$$V_i = U_i - \bigcup\{U_j \mid 0 \leq j \leq i-1\}.$$

It is clear that the first two conditions of the theorem are satisfied; we prove the last part of the theorem.

For $x \in \bigcup\{U_i \mid i \in \mathbb{N}\}$, let $i_x$ be the least $i$ such that $x \in U_i$; clearly, $x \notin U_j$ for $j < i$, so $x \in V_i$. Thus, $\bigcup\{U_i \mid i \in \mathbb{N}\} \subseteq \bigcup\{V_i \mid i \in \mathbb{N}\}$. The reverse inclusion follows immediately from the fact that $V_i \subseteq U_i$ for every $i \in \mathbb{N}$. □

## 6.11 Measures

Measurable spaces provide the natural framework for introducing the notion of measure.

**Definition 6.117.** *Let $(S, \mathcal{E})$ be a measurable space. A* measure *is a function $m : \mathcal{E} \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ that satisfies the following conditions:*
(i) $m(\emptyset) = 0$.
(ii) *For every countable collection $U_0, U_1, \ldots$ of sets in $\mathcal{E}$ that are pairwise disjoint, we have*

$$m\left(\bigcup_{n \in \mathbb{N}} U_n\right) = \sum_{n \in \mathbb{N}} m(U_n)$$

*the* (additivity property).
*We refer to the triple $(S, \mathcal{E}, m)$ as a* measure space.

In particular, if the collection $U_0, U_1, \ldots$ consists of two disjoint sets $U$ and $V$, then

$$m(U \cup V) = m(U) + m(V). \tag{6.3}$$

Observe that if $U, V \in \mathcal{E}$ and $U \subseteq V$, then $V = U \cup (V - U)$, so by the additivity property, $m(V) = m(U) + m(V - U) \geq m(U)$. This shows that $U \subseteq V$ implies $m(U) \leq m(V)$ (the *monotonicity* of measures).

Let $X$ and $Y$ be two subsets of $\mathcal{E}$. Since $X \cup Y = X \cup (Y - X)$, $Y = (Y - X) \cup (Y \cap X)$, and the pairs of sets $X, (Y - X)$ and $(Y - X), (Y \cap X)$ are disjoint, we can write

$$\begin{aligned} m(X \cup Y) &= m(X) + m(Y - X) \\ &= m(X) + m(Y) - m(X \cap Y). \end{aligned} \tag{6.4}$$

The resulting equality

$$m(X \cup Y) + m(X \cap Y) = m(X) + m(Y) \tag{6.5}$$

for $X, Y \in \mathcal{E}$ is known as the *modularity property* of measures.

*Example 6.118.* Let $S$ be a finite set and let $\mathcal{E} = \mathcal{P}(S)$. The mapping $m : \mathcal{P}(S) \longrightarrow \mathbb{R}$ given by $m(U) = |U|$ is a measure on $\mathcal{P}(S)$, as can be verified immediately.

*Example 6.119.* Let $S$ be a set and let $s$ be a fixed element of $S$. Define the mapping $m_s : \mathcal{P}(S) \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ by

$$m_s(U) = \begin{cases} 1 & \text{if } s \in U, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to verify that $m_s$ is a measure defined on $\mathcal{P}(S)$. Indeed, we have $m_s(\emptyset) = 0$. If $U_0, U_1, \ldots$ is a countable collection of pairwise disjoint sets, then $s$ may belong to at most one of these sets. If there is a set $U_i$ such that $s \in U_i$, $s \in \bigcup_{n \in \mathbb{N}} U_i$, so $m_s\left(\bigcup_{n \in \mathbb{N}} U_n\right) = \sum_{n \in \mathbb{N}} m_s(U_n) = 1$. If no such set $U_i$ exists, then $m_s\left(\bigcup_{n \in \mathbb{N}} U_n\right) = \sum_{n \in \mathbb{N}} m_s(U_n) = 0$. In either case, the second condition of Definition 6.117 is satisfied.

The behavior of measures with respect to limits of sequences of sets is discussed next.

**Theorem 6.120.** *Let $(S, \mathcal{E}, m)$ be a measure space. If $(U_0, U_1, \ldots)$ is an increasing or a decreasing sequence of sets from $\mathcal{E}$, then $m(\lim U_n) = \lim m(U_n)$.*

*Proof.* Suppose that $U_0 \subset U_1 \subset \cdots$ is an increasing sequence of sets, so $m(\lim U_n) = m(\bigcup_n U_n)$.

By Theorem 6.116, there exists a sequence $V_0 \subset V_1 \subset \cdots$ of disjoint sets in $\mathcal{E}$ such that $\bigcup U_n = \bigcup V_n$ and $V_0 = U_0$, and $V_n = U_n - V_{n-1}$ for $n \geq 1$. Then,

$$\begin{aligned} m(\lim U_n) &= m\left(\bigcup_n V_n\right) = m(V_0) + \sum_{n \geq 1} m(V_n) \\ &= \lim_{n \to \infty} \left(m(V_0) + \sum_{i=1}^{n} m(V_i)\right) \\ &= \lim_{n \to \infty} m\left(V_0 \cup \bigcup_{n \geq 1} V_i\right) \\ &= \lim_{n \to \infty} m(U_i). \end{aligned}$$

Suppose now that $U_0 \supset U_1 \supset \cdots$ is a decreasing sequence of sets, so $m(\lim U_n) = m(\bigcap_n U_n)$.

Define the sequence of sets $W_0, W_1, \ldots$ by $W_n = U_0 - U_n$ for $n \in \mathbb{N}$. Since this sequence is increasing, we have $m(\bigcup_{n\in\mathbb{N}} W_n) = \lim m(W_n)$ by the first part of the theorem. Thus, we can write

$$m\left(\bigcup_{n\in\mathbb{N}} W_n\right) = \lim m(W_n) = m(U_0) - \lim m(U_n).$$

Since

$$\begin{aligned} m\left(\bigcup_{n\in\mathbb{N}} W_n\right) &= m\left(\bigcup_{n\in\mathbb{N}} (U_0 - U_n)\right) \\ &= m\left(U_0 - \bigcap_{n\in\mathbb{N}} U_n\right) \\ &= m(U_0) - m\left(\bigcap_{n\in\mathbb{N}} U_n\right), \end{aligned}$$

it follows that

$$m(\lim U_n) = m\left(\bigcap_{n\in\mathbb{N}} U_n\right) = \lim m(U_n).$$

□

**Definition 6.121.** *Let $S$ be a set. An* outer measure *on $S$ is a function $\mu : \mathcal{P}(S) \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ that satisfies the following properties:*
(i) $\mu(\emptyset) = 0$.
(ii) *$\mu$ is countably subadditive; that is,*

$$\mu\left(\bigcup_{n\in\mathbb{N}} E_n\right) \leq \sum\{\mu(E_n) \mid n \in \mathbb{N}\}$$

*for every countable family $\{E_n \in \mathcal{P}(S) \mid n \in \mathbb{N}\}$ of subsets of $S$.*
(iii) *$\mu$ is monotonic.*

*A subset $T$ of $S$ is* $\mu$-measurable *if*

$$\mu(H) = \mu(H \cap T) + \mu(H \cap \bar{T})$$

*for every set $H \in \mathcal{P}(S)$.*

**Lemma 6.122.** *Let $S$ be a set and let $\mu$ be an outer measure on a set $S$. A set $T$ is $\mu$-measurable if and only if*

$$\mu(H) \geq \mu(H \cap T) + \mu(H \cap \bar{T})$$

*for every $H \in \mathcal{P}(S)$ such that $\mu(H) < \infty$.*

*Proof.* The necessity of the condition is obvious. Suppose therefore that the condition is satisfied. Since $\mu$ is subadditive, we have

$$\mu(H) \leq \mu(H \cap T) + \mu(H \cap \bar{T}),$$

which implies $\mu(H) = \mu(H \cap T) + \mu(H \cap \bar{T})$. □

**Theorem 6.123.** *Let $\mu$ be an outer measure on a set $S$. The collection of $\mu$-measurable sets is a $\sigma$-field $\mathcal{E}_\mu$ on $S$.*

*Proof.* It is immediate that $\emptyset \in \mathcal{E}_\mu$. Suppose that $T_0, T_1, \ldots$ is a sequence of $\mu$-measurable sets. Then, we can write

$$\mu(H) = \mu(H \cap T_0) + \mu(H \cap \bar{T}_0).$$

By substituting $H \cap T_0$ and $H \cap \bar{T}_0$ by $H$, we obtain

$$\begin{aligned}\mu(H \cap T_0) &= \mu(H \cap T_0 \cap T_1) + \mu(H \cap T_0 \cap \bar{T}_1),\\ \mu(H \cap \bar{T}_0) &= \mu(H \cap \bar{T}_0 \cap T_1) + \mu(H \cap \bar{T}_0 \cap \bar{T}_1),\end{aligned}$$

which implies

$$\mu(H) = \mu(H \cap T_0 \cap T_1) + \mu(H \cap T_0 \cap \bar{T}_1)\mu(H \cap \bar{T}_0 \cap T_1) + \mu(H \cap \bar{T}_0 \cap \bar{T}_1). \quad (6.6)$$

Replacing $H$ by $H \cap (T_0 \cup T_1)$, we obtain the equality

$$\mu(H \cap (T_0 \cup T_1)) = \mu(H \cap T_0 \cap T_1) + \mu(H \cap T_0 \cap \bar{T}_1) + \mu(H \cap \bar{T}_0 \cap T_1). \quad (6.7)$$

Therefore,

$$\mu(H) = \mu(H \cap (T_0 \cup T_1)) + \mu(H \cap \overline{T_0 \cup T_1}),$$

which shows that $T_0 \cup T_1$ is $\mu$-measurable. An easy argument by induction shows that $\bigcup_{i=0}^{n} T_i$ is $\mu$-measurable for every $n \in \mathbb{N}$.

By replacing $H$ in Equality (6.6) by $H \cap \overline{T_0 - T_1} = H \cap (\bar{T}_0 \cup T_1)$, we have

$$\mu(H \cap (\bar{T}_0 \cup T_1)) = \mu(H \cap T_0 \cap T_1) + \mu(H \cap \bar{T}_0 \cap T_1) + \mu(H \cap \bar{T}_0 \cap \bar{T}_1),$$

which allows us to write

$$\mu(H) = \mu(H \cap \overline{T_0 - T_1}) + \mu(H \cap (T_0 \cap -T_1)),$$

which proves that $T_0 - T_1$ is $\mu$-measurable.

If $U_0$ and $U_1$ are two disjoint $\mu$-measurable sets, then Equality (6.7) implies

$$\mu(H \cap (U_0 \cup U_1)) = \mu(H \cap U_0) + \mu(H \cap U_1)$$

for every $H$. Again, an inductive argument allows us to show that if $T_0, \ldots, T_n$ are pairwise disjoint, $\mu$-measurable sets, then

$$\mu\left(H \cap \bigcup_{i=0}^{n} U_i\right) = \sum_{i=0}^{n} \mu(H \cap U_i). \tag{6.8}$$

Define $W_n = \bigcup_{i=0}^{n} T_i$. We have seen that $W_n$ is $\mu$-measurable for every $n \in \mathbb{N}$. Thus, we have

$$\begin{aligned}
\mu(H) &= \mu(H \cap W_n) + \mu(H \cap \bar{W}_n) \\
&= \mu\left(H \cap \left(\bigcup_{i=0}^{n} T_i\right)\right) + \mu(H \cap \bar{W}_n) \\
&\geq \mu\left(H \cap \left(\bigcup_{i=0}^{n} T_i\right)\right) + \mu(H \cap \bar{W}),
\end{aligned}$$

where $W = \bigcup_{i \geq 0} T_i$. By Equality (6.8), we have

$$\mu(H) \geq \sum_{i \geq 0}^{n} \mu(H \cap T_i) + \mu(H \cap \bar{W}) \tag{6.9}$$

for every $n \in \mathbb{N}$. Therefore,

$$\mu(H) \geq \sum_{i \geq 0}^{\infty} \mu(H \cap T_i) + \mu(H \cap \bar{W}),$$

hence $\mu(H) \geq \mu(H \cap W) + \mu(H \cap \bar{W})$. By Lemma 6.122, the set $W$ is $\mu$-measurable. Note also that we have shown that

$$\mu(H) = \sum_{i \geq 0}^{n} \mu(H \cap T_i) + \mu(H \cap \bar{W}) = \mu(H \cap W) + \mu(H \cap \bar{W}). \tag{6.10}$$

Suppose now that the sets $T_0, T_1, \ldots$ are not disjoint. Consider the sequence of pairwise disjoint sets $V_0, V_1, \ldots$ defined by

$$\begin{aligned}
V_0 &= T_0, \\
V_n &= T_n - \bigcup_{i=0}^{n-1} T_i,
\end{aligned}$$

for $n \geq 1$. The measurability of each set $V_n$ is immediate and, by the previous argument, $\bigcup_{n \in \mathbb{N}} V_n$ is $\mu$-measurable. Since $\bigcup_{n \in \mathbb{N}} V_n = \bigcup_{n \in \mathbb{N}} T_n$, it follows that $\bigcup_{n \in \mathbb{N}} T_n$ is $\mu$-measurable. We conclude that the collection of $\mu$-measurable sets is a $\sigma$-field. □

**Corollary 6.124.** *Let $S$ be a set and let $\mu : \mathcal{P}(S) \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ be an outer measure on $S$. The restriction $\mu \restriction_{\mathcal{E}_\mu}$ to the $\sigma$-field $\mathcal{E}_\mu$ is a measure.*

*Proof.* Let $T_0, T_1, \ldots$ be a sequence of sets in $\mathcal{E}_\mu$ that are pairwise disjoint. Choosing $H = W$ in Equality (6.10), we have

$$\mu(W) = \sum_{i \geq 0}^{n} \mu(T_i),$$

which proves that $\mu \restriction_{\mathcal{E}_\mu}$ is indeed a measure. □

**Corollary 6.125.** *Let $\mu$ be an outer measure and let $U_0, U_1, \ldots$ be a sequence of $\mu$-measurable sets. Then, both $\liminf U_n$ and $\limsup U_n$ are $\mu$-measurable sets.*

*Proof.* This statement follows immediately from Theorem 6.123 and from Corollary 6.108. □

Theorem 6.127, which follows gives a technique for constructing outer measures known as *Munroe's Method I* or simply as *Method I* (see [100, 61, 44]).

First, we need the following definition.

**Definition 6.126.** *A* sequential cover *of a set $S$ is a collection $\mathcal{C}$ of subsets of $S$ such that $\emptyset \in \mathcal{C}$, and for every subset $T$ of $S$ there is a countable subcollection $\mathcal{D} = \{D_0, D_1, \ldots\}$ of $\mathcal{C}$ such that $T \subseteq \bigcup_{n=0}^{\infty} D_n$.*

*The family of all* countable *collections of sets from $\mathcal{C}$ that are covers of a set $W \in \mathcal{P}(S)$ is denoted by $\mathfrak{D}_{\mathcal{C},W}$. If the collection $\mathcal{C}$ is clear from the context, the subscript $\mathcal{C}$ will be omitted.*

**Theorem 6.127.** *Let $S$ be a set, $\mathcal{C}$ a sequential cover of the set $S$, and $f : \mathcal{C} \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ a nonnegative function defined on $\mathcal{C}$ such that $f(\emptyset) = 0$.*

*The function $\mu_f : \mathcal{P}(S) \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ given by*

$$\mu_f(T) = \inf \left\{ \sum_{U \in \mathcal{D}} f(U) \mid \mathcal{D} \in \mathfrak{D}_{\mathcal{C},T} \right\}$$

*for $T \in \mathcal{P}(S)$ is an outer measure on $S$.*

*Proof.* Since $\emptyset$ is covered by the empty collection, and an empty sum has the value 0, it follows that $\mu_f(\emptyset) = 0$.

If $T, T' \in \mathcal{P}(S)$ and $T \subseteq T'$, then any cover of $T'$ is also a cover of $T$; that is, $\mathfrak{D}_{\mathcal{C},T'} \subseteq \mathfrak{D}_{\mathcal{C},T}$. Therefore, $\mu_f(T) \leq \mu_f(T')$.

Let $\{T_n \mid n \in \mathbb{N}\}$ be a countable collection of subsets of $S$. If $\mu_f(T_n) = +\infty$ for one of the members of this collection, then the subadditivity of $\mu_f$,

$$\mu_f \left( \bigcup_{n \in \mathbb{N}} U_n \right) \leq \sum_{n \in \mathbb{N}} \mu_f(U_n),$$

is satisfied. Therefore, we assume now that the value $\mu_f(T_n)$ is finite for each $n \in \mathbb{N}$.

The definition of $\mu_f(T_n)$ as an infimum allows us to assume the existence of a collection of sets $\mathcal{D}_n \in \mathfrak{D}_{U_n}$ such that

$$\sum_{U \in \mathcal{D}_n} f(U) \leq \mu_f(T_n) + \frac{\epsilon}{2^n}.$$

Consider the collection $\mathcal{D} = \bigcup_{n \in \mathbb{N}} \mathcal{D}_n$. $\mathcal{D}$ is a cover for $\bigcup_{n \in \mathbb{N}} T_n$. Therefore, by the definition of $\mu_f$, we have

$$\begin{aligned}
\mu_f\left(\bigcup_{n \in \mathbb{N}} T_n\right) &\leq \sum\{f(U) \mid U \in \mathcal{D}\} \\
&\leq \sum_{n \in \mathbb{N}} \sum_{U \in \mathcal{D}_n} f(U) \\
&\leq \sum_{n \in \mathbb{N}} \mu_f(T_n) + \epsilon \sum_{n \in \mathbb{N}} \frac{1}{2^n} \\
&= \sum_{n \in \mathbb{N}} \mu_f(T_n) + 2\epsilon.
\end{aligned}$$

Since this inequality holds for every $\epsilon$, it follows that

$$\mu_f\left(\bigcup_{n \in \mathbb{N}} T_n\right) \leq \sum\{\mu_f(T_n) \mid n \in \mathbb{N}\},$$

which proves that $\mu_f$ is subadditive. We conclude that $\mu_f$ is an outer measure. □

**Corollary 6.128.** *Let $S$ be a set, $\mathcal{C}$ a sequential cover of the set $S$, and $f : \mathcal{C} \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ a function such that $f(\emptyset) = 0$. The outer measure $\mu_f$ is the unique outer measure on $S$ that satisfies the following properties:*

(i) *$\mu_f(U) \leq f(U)$ for every $U \in \mathcal{C}$, and*
(ii) *if $\mu'$ is an outer measure such that $\mu'(U) \leq f(U)$ for every $U \in \mathcal{C}$ then $\mu'(T) \leq \mu_f(T)$ for every $T \in \mathcal{P}(S)$.*

*Proof.* Since $\{\emptyset, U\}$ is a cover for $U$, the inequality $\mu_f(U) \leq f(U)$ is immediate for every $U \in \mathcal{C}$.

Let $\mu'$ be an outer measure such that $\mu'(U) \leq f(U)$ for every $U \in \mathcal{C}$ and let $\mathcal{D}$ be a sequential cover of a set $T \in \mathcal{P}(S)$. Then, we have

$$\mu'(T) \leq \mu'\left(\bigcup \mathcal{D}\right) \leq \{\mu'(U) \mid U \in \mathcal{D}\} \leq \sum\{f(U) \mid U \in \mathcal{D}\}$$

so $\mu'(T) \leq \mu_f(T)$. The uniqueness of $\mu_f$ follows by changing the roles of $\mu_f$ and $\mu'$. □

**Corollary 6.129.** *Let $S$ be a set, $\mathcal{C}'$ and $\mathcal{C}$ two sequential covers of $S$ such that $\mathcal{C}' \subseteq \mathcal{C}$, and $f : \mathcal{C} \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ a function such that $f(\emptyset) = 0$. If $\mu'_f$ and $\mu_f$ are the outer measures that correspond to the collections $\mathcal{C}'$ and $\mathcal{C}$, respectively, then $\mu_f(T) \leq \mu'_f(T)$ for $T \in \mathcal{P}(S)$.*

*Proof.* Observe that if $\mathfrak{D}'_T \subseteq \mathfrak{D}_T$, where $\mathfrak{D}'_T$ and $\mathfrak{D}_T$ are the families of countable collections of sets from $\mathcal{C}'$ and $\mathcal{C}$, respectively, that are covers of $T$, then the definitions of $\mu'_f$ and $\mu_f$ immediately imply the desired inequality. □

*Example 6.130.* Theorem 6.127 allows us to introduce a very important outer measure on $\mathbb{R}$. Let $\mathcal{C}$ be the collection of open intervals of $\mathbb{R}$ to which the empty set is added.

Define the function $f : \mathcal{C} \longrightarrow \mathbb{R}$ by $f(a, b) = b - a$ for every open interval $(a, b) \in \mathcal{C}$ and $f(\emptyset) = 0$. For a subset $T$ of $\mathbb{R}$, the value of the outer measure $\mu(T) = m_f(T)$ is

$$\mu(T) = \inf\left\{\sum_{n\in\mathbb{N}}(b_n - a_n) \in \mathcal{C} \mid T \subseteq \bigcup_n (a_n, b_n)\right\},$$

where the infimum is considered over all countable collections of open intervals $(a_n, b_n)$ that cover the set $T$. This is the *Lebesgue outer measure* of the set $T$.

Let $\mu$ be the Lebesgue outer measure on $\mathbb{R}$. We have $\mu([a, b]) = b - a$. Since $[a, b] \subseteq (a - \epsilon, b + \epsilon)$ for every $\epsilon > 0$, it follows that $\mu([a, b]) < b - a + 2\epsilon$ for every $\epsilon > 0$, so $\mu([a, b]) \leq b - a$. On the other hand, $(a, b) \subseteq [a, b]$, so $\mu([a, b]) \geq b - a$, which yields $\mu([a, b]) = b - a$.

This type of measure can be generalized to $\mathbb{R}^n$ by defining $\mathcal{C}$ as the collection of $n$-dimensional intervals of the form $I = (a_1, b_1) \times \cdots \times (a_n, b_n)$ to which we add the empty set and letting $f(I)$ be the volume $vol(I) = \prod_{i=1}^n |b_i - a_i|$ of $I$. Thus, the Lebesgue measure of a set $T \subseteq \mathbb{R}^n$ is

$$\mu(T) = \inf\left\{\sum vol(I) \mid I \in \mathcal{C}, T \subseteq \bigcup I\right\}, \tag{6.11}$$

**Definition 6.131.** *An outer measure $\mu$ on a set $S$ is* regular *if for every $T \in \mathcal{P}(S)$ there exists a $\mu$-measurable set $U$ such that $T \subseteq U$ and $\mu(T) = \mu(U)$.*

*Example 6.132.* Let $\mu$ be the Lebesgue outer measure on $\mathbb{R}^n$ and let $T$ be a subset of $\mathbb{R}^n$. For every $m \in \mathbb{N}$ there exists a countable collection of intervals $\{I_m^k \mid k \in \mathbb{N}\}$ such that

$$\mu(T) \leq \sum_{k\in\mathbb{N}} \mu(I_m^k) < \mu(T) + \frac{1}{m} < \mu\left(\bigcup_{k\in\mathbb{N}} I_m^k\right) + \frac{1}{m}.$$

Let $U = \bigcap_{m\in\mathbb{N}} \bigcap_{k\in\mathbb{N}} I_m^k$. Clearly, $U$ is $\mu$-measurable and $T \subseteq U$, so $\mu(T) \leq \mu(U)$. Since $U \subseteq \bigcap_{k\in\mathbb{N}} I_m^k$, we have

$$\mu(U) \leq \mu\left(\bigcap_{k\in\mathbb{N}} I_m^k\right) \leq \sum_{k\in\mathbb{N}} \mu\left(\bigcap_{k\in\mathbb{N}} I_m^k\right) \leq \mu(T) + \frac{1}{m},$$

so $\mu(U) \leq \mu(T)$. Consequently, $\mu(U) = \mu(T)$, which proves that the Lebesgue outer measure on $\mathbb{R}^n$ is regular.

**Theorem 6.133.** *Let $S$ be a set and let $(S_0, S_1, \ldots)$ be a sequence of subsets of $S$. If $\mu$ is a regular outer measure on $S$, then $\mu(\liminf_n S_n) \leq \liminf_n \mu(S_n)$.*

*Proof.* Since $\mu$ is regular, for each $n \in \mathbb{N}$ there exists a $\mu$-measurable set $U_n$ such that $S_n \subseteq U_n$ and $\mu(S_n) = \mu(U_n)$. Then, $\liminf_n \mu(S_n) \liminf_n \mu(U_n)$. Since $\liminf_n \mu(U_n)$ is measurable (by Corollary 6.125), we have

$$\mu(\liminf_n S_n) \leq \mu(\liminf_n U_n) \leq \liminf_n \mu(U_n) = \liminf_n \mu(S_n).$$

□

**Corollary 6.134.** *Let $\mu$ be an outer measure on a set $S$. If $\boldsymbol{S} = (S_0, S_1, \ldots)$ is an expanding sequence of subsets of $S$, then $\mu(\lim_n S_n) = \lim_n \mu(S_n)$.*

*Proof.* Since $\mathbf{S}$ is an expanding sequence $\lim_n S_n = \bigcup_n S_n$, then $\mu(\lim_n S_n) \geq \mu(S_n)$ for $n \in \mathbb{N}$, so $\mu(\lim_n S_n) \geq \lim_n \mu(S_n)$. On the other hand, Theorem 6.133 implies $\mu(\lim S_n) \leq \lim \mu(S_n)$, which gives the desired equality. □

For finite regular outer measures, the measurability condition can be simplified, as shown next.

**Theorem 6.135.** *Let $\mu$ be a regular outer measure on a set $S$ such that $\mu(S)$ is finite. A subset $T$ of $S$ is measurable if and only if $\mu(S) = \mu(T) + \mu(\bar{T})$, where $\bar{T} = S - T$.*

*Proof.* The condition is clearly necessary. To prove its sufficiency, let $T$ be a subset of $S$ such that $\mu(S) = \mu(T) + \mu(\bar{T})$. By Lemma 6.122, to prove that $T$ is measurable, it suffices to show that if $H$ is a set with $\mu(H) < \infty$, then $\mu(H) \geq \mu(H \cap T) + \mu(H \cap \bar{T})$.

The regularity of $\mu$ implies the existence of a $\mu$-measurable set $K$ such that $H \subseteq K$ and $\mu(H) = \mu(K)$. Since $K$ is measurable, we have

$$\begin{aligned}\mu(H) &= \mu(H \cap K) + \mu(H \cap \bar{K}),\\ \mu(\bar{H}) &= \mu(\bar{H} \cap K) + \mu(\bar{H} \cap \bar{K}).\end{aligned}$$

This implies

$$\begin{aligned}\mu(S) &= \mu(T) + \mu(\bar{T})\\ &= \mu(T \cap K) + \mu(T \cap \bar{K}) + \mu(\bar{T} \cap K) + \mu(\bar{T} \cap \bar{K})\\ &\geq \mu(K) + \mu(\bar{K}) = \mu(S).\end{aligned}$$

Thus,

$$\mu(T \cap K) + \mu(T \cap \bar{K}) + \mu(\bar{T} \cap K) + \mu(\bar{T} \cap \bar{K}) = \mu(K) + \mu(\bar{K}) = \mu(S).$$

Since $\mu(\bar{K}) \leq \mu(T \cap \bar{K}) + \mu(\bar{T} \cap \bar{K})$, it follows that $\mu(K \cap T) + \mu(K \cap \bar{T}) \leq \mu(K)$. Since $H \cap T \subseteq K \cap T$ and $H \cap \bar{T} \subseteq K \cap \bar{T}$, we have $\mu(H \cap T) + \mu(H \cap \bar{T}) \leq \mu(K) = \mu(H)$, which shows that $T$ is indeed $\mu$-measurable. □

## Exercises and Supplements

1. Prove that the family of subsets $\{(-n, n) \mid n \in \mathbb{N}\} \cup \{\emptyset, \mathbb{R}\}$ is a topology on $\mathbb{R}$.
2. Let $S$ be a set and let $s_0$ be an element of $S$. Prove that the family of subsets $\mathcal{O}_{s_0} = \{L \in \mathcal{P}(S) \mid s_0 \in L\} \cup \{\emptyset\}$ is a topology on $S$.
3. Let $(S, \mathcal{O})$ be a topological space, $L$ be an open set in $(S, \mathcal{O})$, and $H$ be a closed set.
   a) Prove that a set $V$ is open in the subspace $(L, \mathcal{O} \restriction_L)$ if and only if $V$ is open in $(S, \mathcal{O})$ and $V \subseteq L$.
   b) Prove that a set $W$ is closed in the subspace $(H, \mathcal{O} \restriction_H)$ if and only if $W$ is closed in $(S, \mathcal{O})$ and $W \subseteq H$.
4. Let $(S, \mathcal{O})$ be a topological space where $\mathcal{O} = \{\emptyset, U, V, S\}$, where $U$ and $V$ are two subsets of $S$. Prove that either $\{U, V\}$ is a partition of $S$ or one of the sets $\{U, V\}$ is included in the other.
5. Let $(S, \mathcal{O})$ be a topological space and let $\mathbf{I}$ be its interior operator. Prove that the poset of open sets $(\mathcal{O}, \subseteq)$ is a complete lattice, where $\sup \mathcal{L} = \bigcup \mathcal{L}$ and $\inf \mathcal{L} = \mathbf{I}(\bigcap \mathcal{L})$ for every family of open sets $\mathcal{L}$.
6. Let $(S, \mathcal{O})$ be a topological space, let $\mathbf{K}$ be its interior operator and let $\mathcal{K}$ be its collection of closed sets. Prove that $(\mathcal{K}, \subseteq)$ is a complete lattice, where $\sup \mathcal{L} = \mathbf{K}(\bigcup \mathcal{L})$ and $\inf \mathcal{L} = \bigcap \mathcal{L}$ for every family of closed sets.
7. Prove that if $U, V$ are two subsets of a topological space $(S, \mathcal{O})$, then $\mathbf{K}(U \cap V) \subseteq \mathbf{K}(U) \cap \mathbf{K}(V)$. Formulate an example where this inclusion is strict.
8. Let $T$ be a subspace of the topological space $(S, \mathcal{O})$. Let $\mathbf{K}_S, \mathbf{I}_S$, and $\partial_S$ be the closure, interior and border operators associated to $S$ and $\mathbf{K}_T, \mathbf{I}_T$ and $\partial_T$ the corresponding operators associated to $T$. Prove that
   a) $\mathbf{K}_T(U) = \mathbf{K}_S(U) \cap T$,
   b) $\mathbf{I}_S(U) \subseteq \mathbf{I}_T(U)$, and
   c) $\partial_T U \subseteq \partial_S U$

   for every subset $U$ of $T$.
9. Let $(S, \mathcal{O})$ be a topological space and let $\mathbf{K}$ and $\mathbf{I}$ be its associated closure and interior operator, respectively. Define the mappings $\phi, \psi : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ by $\phi(U) = \mathbf{I}(\mathbf{K}(U))$ and $\psi(U) = \mathbf{K}(\mathbf{I}(U))$ for $U \in \mathcal{P}(S)$.
   a) Prove that $\phi(U)$ is an open set and $\psi(U)$ is a closed set for every set $U \in \mathcal{P}(S)$.

b) Prove that $\psi(H) \subseteq H$ for every closed set $H$ and $L \subseteq \phi(L)$ for every open set $L$.
c) Prove that $\phi(\phi(U)) = \phi(U)$ and $\psi(\psi(U)) = \psi(U)$ for every $U \in \mathcal{P}(S)$.
d) Let $(\mathbf{J}_1, \ldots, \mathbf{J}_n)$ be a sequence such that $\mathbf{J}_i \in \{\mathbf{K}, \mathbf{I}\}$. Prove that there are at most seven distinct sets of the form $\mathbf{J}_n(\cdots(\mathbf{J}_1(U))\cdots)$ for every set $U \in \mathcal{P}(S)$, and give an example of a topological space $(S, \mathcal{O})$ and a subset $U$ of $S$ such that these seven sets are pairwise distinct.

10. Let $\mathcal{S}$ be the set of subsets of $\mathbb{R}$ such that, for every $U \in \mathcal{S}$, $x \in U$ implies $-x \in U$. Prove that $\{\emptyset\} \cup \mathcal{S}$ is a topology on $\mathbb{R}$.
11. Let $(S, \mathcal{O})$ be a topological space, and $U$ and $U'$ be two subsets of $S$.
    a) Prove that $\partial(U \cup V) \subseteq \partial U \cup \partial V$.
    b) Prove that $\partial U = \partial(S - U)$.
12. Let $(S, \mathcal{O})$ be a topological space. The subsets $X$ and $Y$ are said to be *separated* if $X \cap \mathbf{K}(Y) = \mathbf{K}(X) \cap Y = \emptyset$.
    a) Prove that $X$ and $Y$ are separated sets in $(S, \mathcal{O})$ if and only if they are disjoint and clopen in the subspace $X \cup Y$.
    b) Prove that two disjoint open sets or two disjoint closed sets in $(S, \mathcal{O})$ are separated.
13. Let $\mathcal{B}$ be a base for a topological space $(S, \mathcal{O})$. Prove that if $\mathcal{B}'$ is a collection of subsets of $S$ such that $\mathcal{B} \subseteq \mathcal{B}' \subseteq \mathcal{O}$, then $\mathcal{B}'$ is a basis for $\mathcal{O}$.
14. Let $(S, \mathcal{O})$ be a topological space, $U$ and $U'$ two subsets of $S$, and $\mathcal{B}$ and $\mathcal{B}'$ two bases in the subspaces $(U, \mathcal{O}\restriction_U)$ and $(U', \mathcal{O}\restriction_{U'})$, respectively. Prove that $\mathcal{B} \vee \mathcal{B}'$ is a basis in the subspace $U \cup U'$.

    **Solution:** Let $M$ be an open set in the subspace $U \cup V$. By the definition of the subspace topology, there exists an open set $L \in \mathcal{O}$ such that $M = L \cap (U \cup V) = (L \cap U) \cup (L \cap V)$, so $L$ is the union of two open sets, $L \cap U$ and $L \cap U'$, in the subspaces $U$ and $U'$. Since $\mathcal{B}$ is a basis in $U$, there is a subcollection $\mathcal{B}_1$ such that $L \cap U = \bigcup \mathcal{B}_1$. Similarly, $\mathcal{B}'$ contains a subcollection $\mathcal{B}'_1$ such that $L \cap U = \bigcup \mathcal{B}'_1$. Therefore, $M = \bigcup \mathcal{B}_1 \cup \bigcup \mathcal{B}'_1 = \bigcup \mathcal{B}_1 \vee \mathcal{B}'_1$.
15. Let $S$ be an uncountable set and let $(S, \mathcal{O})$ be the cofinite topology on $S$.
    a) Prove that every nonfinite set is dense.
    b) Prove that there is no countable basis for this topological space. What does this say about Theorem 6.46?
16. Let $\mathcal{C}$ be the family of open intervals $\mathcal{C} = \{(a, b) \mid a, b \in \mathbb{R} \text{ and } ab > 0\}$. Prove that:
    a) Every open set $L$ of $(\mathbb{R}, \mathcal{O})$ contains a member of $\mathcal{C}$.
    b) $\mathcal{C}$ is not a basis for the topology $\mathcal{O}$.
17. Let $\mathcal{C}$ be a chain of subsets of a set $S$ such that $\bigcup \mathcal{C} = S$. Prove that $\mathcal{C}$ is the basis of a topology.
18. Prove that if $(S, \mathcal{O})$ is a topological space such that $\mathcal{O}$ is finite, then $(S, \mathcal{O})$ is compact.
19. Prove that the topological space $(\mathbb{R}, \mathcal{O})$ introduced in Example 6.4 is not compact.

20. Let $(S, \mathcal{O})$ be a compact space and let $\mathbf{H} = (H_0, H_1, \ldots)$ be a non-increasing sequence of nonempty and closed subsets of $S$. Prove that $\bigcap_{i \in \mathbb{N}} H_i$ is nonempty.
21. Let $(S_1, \mathcal{O}_1)$ and $(S_2, \mathcal{O}_2)$ be two topological spaces and let $f : S_1 \longrightarrow S_2$ be a continuous surjective function. Prove that if $(S_2, \mathcal{O}_2)$ is compact, then $(S_1, \mathcal{O}_1)$ is compact.
22. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a continuous function defined on the topological space $(\mathbb{R}, \mathcal{O})$. Prove that if $f(q) = 0$ for every $q \in \mathbb{Q}$, then $f(x) = 0$ for every $x \in \mathbb{R}$.
23. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a continuous function in $x_0$. Prove that if $f(x_0) > 0$, then there exists an open interval $(a, b)$ such that $x_0 \in (a, b)$ and $f(x) > 0$ for every $x \in (a, b)$.
24. Let $(S, \mathcal{O}_{s_0})$ be the topological space defined in Exercise 2, where $s_0 \in S$. Prove that any continuous function $f : S \longrightarrow \mathbb{R}$ is a constant function.
25. Let $(S, \mathcal{O})$ and $(T, \mathcal{O}')$ be two topological spaces and let $\mathcal{B}'$ be a basis of $(T, \mathcal{O}')$. Prove that $f : S \longrightarrow T$ is continuous if and only if $f^{-1}(B) \in \mathcal{O}$ for every $B \in \mathcal{B}'$.

Let $(S_1, \mathcal{O}_1)$ and $(S_2, \mathcal{O}_2)$ be two topological spaces and let $f : S_1 \longrightarrow S_2$ be a function. Then $f$ is an *open function* if $f(L)$ is open for every open set $L$, where $L \in \mathcal{O}_1$; the function $f$ is a *closed function* if $f(H)$ is closed for every closed set $H$ in $S_1$.

26. Let $(S_1, \mathcal{O}_1)$ and $(S_2, \mathcal{O}_2)$ be two topological spaces and let $\mathbf{K}_i$ and $\mathbf{I}_i$ be the closure and interior operators of the space $S_i$ for $i = 1, 2$.
    a) Prove that $f : S_1 \longrightarrow S_2$ is an open function if and only if $f(\mathbf{I}_1(U)) \subseteq \mathbf{I}_2(f(U))$ for every $U \in \mathcal{P}(S_1)$.
    b) Prove that $f : S_1 \longrightarrow S_2$ is a closed function if and only if $\mathbf{K}_2(f(U)) \subseteq f(\mathbf{K}_1(U))$ for every $U \in \mathcal{P}(S_1)$.
    c) Prove that a bijection $f : S_1 \longrightarrow S_2$ is open if and only if it is closed.
27. Prove that the function $f : \mathbb{R} \longrightarrow \mathbb{R}$ defined by $f(x) = x^2$ for $x \in \mathbb{R}$ is continuous but not open.
28. Prove that if $a < b$ and $c < d$, then the subspaces $[a, b]$ and $[c, d]$ are homeomorphic.
29. Let $(S, \mathcal{O})$ be a connected topological space and $f : S \longrightarrow \mathbb{R}$ be a continuous function. Prove that if $x, y \in S$, then for every $r \in [f(x), f(y)]$ there is $z \in S$ such that $f(z) = r$.
30. Let $a$ and $b$ be two real numbers such that $a \leq b$. Prove that if $f : [a, b] \longrightarrow [a, b]$ is a continuous function, then there is $c \in [a, b]$ such that $f(c) = c$.
31. Prove that a topological space $(S, \mathcal{O})$ is connected if and only if $\partial T = \emptyset$ implies $T \in \{\emptyset, S\}$ for every $T \in \mathcal{P}(S)$.

Let $(S, \mathcal{O})$ be a topological space and let $x$ and $y$ be two elements of $S$. A *continuous path* between $x$ and $y$ is a continuous function $f : [0, 1] \longrightarrow S$

such that $f(0) = x$ and $f(1) = y$. We refer to $x$ as the *origin* and to $y$ as the *destination* of $f$.

$(S, \mathcal{O})$ is said to be *arcwise connected* if any two points $x$ and $y$ are the origin and destination of a continuous path.

32. Prove that any arcwise connected topological space is connected.
33. Let $(S, \mathcal{O})$ be a $T_0$ topological space. Define the relation "$\leq$" on $S$ by $x \leq y$ if $x \in \mathbf{K}(\{y\})$. Prove that $\leq$ is a partial order.
34. Let $(S, \mathcal{O})$ be a $T_4$ topological space.
    a) Let $H$ and $H'$ be two closed sets and $L$ be an open set such that $H \cap H' \subseteq L$. Prove that there exists two open sets $U$ and $U'$ such that $H \subseteq U$, $H' \subseteq U'$, and $L = U \cap U'$.
    b) If $\{H_1, \ldots, H_p\}$ is a collection of closed sets such that $p \geq 2$ and $\bigcap_{i=1}^{p} H_i = \emptyset$, prove that there exists a family of open sets $\{U_1, \ldots, U_p\}$ such that $\bigcap_{i=1}^{p} U_i = \emptyset$ and $H_i \subseteq U_i$ for $1 \leq i \leq p$.

    **Solution:** Observe that the sets $H - L$ and $H' - L$ are closed and disjoint sets. Since $(S, \mathcal{O})$ is $T_4$, there are two disjoint open sets $V$ and $V'$ such that $H - L \subseteq V$ and $H' - L \subseteq V'$. Define the open sets $U$ and $U'$ as $U = V \cup L$ and $U' = V' \cup L$. It is clear that $U$ and $U'$ satisfy the requirements of the statement.

    The second part is an extension of Definition 6.86. The argument is by induction on $p$. The base case, $p = 2$, follows immediately from the definition of $T_4$ spaces.

    Suppose that the statement holds for $p$, and let $\{H_1, \ldots, H_{p+1}\}$ be a collection of closed sets such that $\bigcap_{i=1}^{p+1} H_i = \emptyset$.

    By applying the inductive hypothesis to the collection of $p$ closed sets $\{H_1, \ldots, H_{p-1}, H_p \cap H_{p+1}\}$, we obtain the existence of the open sets $U_1, \ldots, U_{p-1}, U$ such that $H_i \subseteq U_i$ for $1 \leq i \leq p-1$, $H_p \cap H_{p+1} \subseteq U$, and $\left(\bigcap_{j=1}^{p-1} U_j\right) \cap U = \emptyset$. By the first part of this supplement, we obtain the existence of two open sets $U_p$ and $U_{p+1}$ such that $H_p \subseteq U_p$, $H_{p+1} \subseteq U_{p+1}$, and $U = U_p \cap U_{p+1}$. Note that $\bigcap_{j=1} U_j = \emptyset$, which concludes the argument.
35. Let $(S, \mathcal{O})$ be a $T_4$ topological space and let $\mathfrak{L} = \{L_1, \ldots, L_p\}$ be an open cover of $S$.
    a) Prove that for every $k$, $1 \leq k \leq p$ there exist $k$ open sets $V_1, \ldots, V_k$ such that the collection $\{S - \mathbf{K}(V_1), \ldots, S - \mathbf{K}(V_k), L_{k+1}, \ldots, L_p\}$ is an open cover of $S$ and for the closed sets $H_j = S - V_j$ we have $H_j \subseteq L_j$ for $1 \leq j \leq k$.
    b) Conclude that for every open cover $\mathfrak{L} = \{L_1, \ldots, L_p\}$ of $S$ there is a closed cover $\mathcal{H} = \{H_1, \ldots, H_p\}$ of $S$ such that $H_i \subseteq L_i$ for $1 \leq i \leq p$.

    **Solution:** The proof of the first part is by induction on $k$, $1 \leq k \leq p$. For the base case, $k = 1$, observe that $S - L_1 \subseteq \bigcup_{j=2}^{p} L_j$ because $\mathfrak{L}$ is a cover. Since $(S, \mathcal{O})$ is a $T_4$ space, there exists an open set $V_1$ such that

$S - L_1 \subseteq V_1 \subseteq \mathbf{K}(V_1) \subseteq \bigcup_{j=2}^{p} L_j$. For $H_1 = S - V_1$, it is clear that $H_1 \subseteq L_1$ and $\{S - \mathbf{K}(V_1), L_2, \ldots, L_p\}$ is an open cover of $S$.

Suppose that the statement holds for $k$. This implies

$$S - L_{k+1} \subseteq \bigcup_{j=1}^{k} (S - \mathbf{K}(V_j)) \cup \bigcup_{j=k+2}^{p} L_j.$$

Again, by the property of $T_4$ spaces, there is an open set $V_{k+1}$ such that

$$S - L_{k+1} \subseteq V_{k+1} \subseteq \mathbf{K}(V_{k+1}) \bigcup_{j=1}^{k} (S - \mathbf{K}(V_j)) \cup \bigcup_{j=k+2}^{p} L_j.$$

Thus, $\{S - \mathbf{K}(V_1), \ldots, S - \mathbf{K}(V_k), S - \mathbf{K}(V_{K+1}), L_{k+2}, \ldots, L_p\}$ is an open cover of $S$ and $H_{k+1} = S - V_{k+1} \subseteq L_{k+1}$, which concludes the inductive step.

The second part follows immediately from the first by taking $k = p$. Indeed, since $\{S - \mathbf{K}(V_1), \ldots, S - \mathbf{K}(V_p)\}$ is a cover of $S$ and $S - \mathbf{K}(V_i) \subseteq H_i$ for $1 \leq i \leq p$, it follows immediately that $\mathcal{H}$ is a cover of $S$.

36. Let $(S, \mathcal{O})$ be a $T_4$ topological space, $\mathcal{L} = \{L_1, \ldots, L_p\}$ be an open cover of $S$, and $\mathcal{H} = \{H_1, \ldots, H_p\}$ be a closed cover of $S$ such that $H_i \subseteq L_i$ for $1 \leq i \leq p$ and $\bigcap \mathcal{H} = \emptyset$.
    a) Prove that for every $k$, $1 \leq k \leq p$ there exist $k$ open sets $M_1, \ldots, M_k$ such that:
        i. $H_j \subseteq M_j$ and $\mathbf{K}(M_j) \subseteq L_j$ for $1 \leq j \leq k$,
        ii. the collection $\{M_1, \ldots, M_k, L_{k+1}, \ldots, L_p\}$ is an open cover of $S$, and
        iii. $\bigcap_{i=1}^{k} \mathbf{K}(M_i) \cap \bigcap_{i=k+1}^{p} H_i = \emptyset$.
    b) Prove that there exists an open cover $\mathcal{M} = \{M_1, \ldots, M_p\}$ of $S$ such that $M_i \subseteq L_i$ for $1 \leq i \leq p$ and $\bigcap \mathcal{M} = \emptyset$.

    **Solution:** The proof of the first part is by induction on $k$, $1 \leq k \leq p$. For the base case, $k = 1$, observe that $H_1 \cap \bigcap_{i=2}^{p} H_i = \emptyset$ implies $H_1 \subseteq S - \bigcap_{i=2}^{p} H_i$, so $H_1 \subseteq L_1 \cap (S - \bigcap_{i=2}^{p} H_i)$. This implies the existence of an open set $M_1$ such that

$$H_1 \subseteq M_1 \subseteq \mathbf{K}(M_1) \subseteq L_1 \cap \left( S - \bigcap_{i=2}^{p} H_i \right),$$

which implies $\mathbf{K}(M_1) \subseteq L_1$ and $\mathbf{K}(M_1) \cap \bigcap_{i=2}^{p} H_i = \emptyset$.

Suppose that the statement holds for $k$. We have $H_{k+1} \subseteq L_{k+1}$ and, by the inductive hypothesis, $H_{k+1} \subseteq S - \left( \bigcap_{i=1}^{k} \mathbf{K}(M_i) \cap \bigcap_{i=k+2}^{p} H_i \right)$. Thus,

$$H_{k+1} \subseteq L_{k+1} \cap \left( S - \left( \bigcap_{i=1}^{k} \mathbf{K}(M_i) \cap \bigcap_{i=k+2}^{p} H_i \right) \right).$$

By the $T_4$ separation property, there exists an open set $M_{k+1}$ such that

$$H_{k+1} \subseteq M_{k+1} \subseteq \mathbf{K}(M_{k+1}) \subseteq L_{k+1} \cap \left( S - \left( \bigcap_{i=1}^{k} \mathbf{K}(M_i) \cap \bigcap_{i=k+2}^{p} H_i \right) \right),$$

which implies $\mathbf{K}(M_{k+1}) \subseteq L_{k+1}$ and $\bigcap_{i=1}^{k+1} \mathbf{K}(M_i) \cap \bigcap_{i=k+2}^{p} H_i = \emptyset$.

The second part of the supplement follows directly from the first part.

37. Prove that if $(S, \mathcal{P}(S))$ and $(S', \mathcal{P}(S'))$ are two discrete topological spaces, then their product is a discrete topological space.
38. Let $(S, \mathcal{O}), (S, \mathcal{O}')$ be two topological spaces. Prove that the collection

$$\{S \times L' \mid L' \in \mathcal{O}'\} \cup \{L \times S' \mid L \in \mathcal{O}\}$$

is a subbase for the product topology $\mathcal{O} \times \mathcal{O}'$.

39. Let $(S, \mathcal{O}), (S, \mathcal{O}')$ be two topological spaces and let $(S \times S', \mathcal{O} \times \mathcal{O}')$ be their product.
    a) Prove that for all sets $T, T'$ such that $T \subseteq S$ and $T' \subseteq S'$, $\mathbf{K}(T \times T') = \mathbf{K}(T) \times \mathbf{K}(T')$ and $\mathbf{I}(T \times T') = \mathbf{I}(T) \times \mathbf{I}(T')$.
    b) Prove that $\partial(T \times T') = (\partial(T) \times \mathbf{k}(T')) \cup (\mathbf{k}(T) \times \partial T')$.
40. Prove that the following classes of topological spaces are closed with respect to the product of topological spaces:
    a) the class of spaces that satisfy the first axiom of countability;
    b) the class of spaces that satisfy the second axiom of countability;
    c) the class of separable spaces.
41. Prove that, for a topological space $(S, \mathcal{O})$, the following statements are equivalent:
    a) $(S, \mathcal{O})$ is connected.
    b) If $S = L_1 \cup L_2$ and $L_1 \cap L_2 = \emptyset$, where $L_1$ and $L_2$ are open, then $L_1 = \emptyset$ or $L_2 = \emptyset$.
    c) If $S = H_1 \cup H_2$ and $H_1 \cap H_2 = \emptyset$, where $H_1$ and $H_2$ are closed, then $H_1 = \emptyset$ or $H_2 = \emptyset$.
    d) If $K$ is a clopen set, then $K = \emptyset$ or $K = S$.
42. Prove that any subspace of a totally disconnected topological space is totally disconnected, and prove that a product of totally disconnected topological spaces is totally disconnected.
43. Let $S$ be a set and let $\mathcal{C}$ be a collection of subsets of $S$. Define the collections of sets

$$\begin{aligned} \mathcal{C}' &= \mathcal{C} \cup \{S - T \mid T \in \mathcal{C}\}, \\ \mathcal{C}'' &= \left\{ \bigcap \mathcal{D} \mid \mathcal{D} \subseteq \mathcal{C}' \right\}, \\ \mathcal{C}''' &= \left\{ \bigcup \mathcal{D} \mid \mathcal{D} \subseteq \mathcal{C}'' \right\}. \end{aligned}$$

Prove that $\mathcal{C}'''$ equals the $\sigma$-field generated by $\mathcal{C}$.

44. Let $S$ and $T$ be two sets and let $f : S \longrightarrow T$ be a function. Prove that if $\mathcal{E}'$ is a $\sigma$-field on $T$, then $\{f^{-1}(V) \mid V \in \mathcal{E}'\}$ is a $\sigma$-field on $A$.
45. Prove that any $\sigma$-field $\mathcal{E}$ contains the empty set; further, prove that if $\mathbf{s} = (S_0, S_1, \ldots)$ is a sequence of sets of $\mathcal{E}$, then both $\liminf \mathbf{s}$ and $\limsup \mathbf{s}$ belong to $\mathcal{E}$.
46. Let $S$ be an infinite set and let $\mathcal{E}$ be the collection $\mathcal{E} = \{E \in \mathcal{P}(S) \mid E \text{ is finite or cofinite}\}$. Prove that $\mathcal{E}$ is a field of sets on $X$ but not a $\sigma$-field.
47. Let $S$ be a set and let $\mathcal{E}$ be a $\sigma$-field. Define the function $m : \mathcal{E} \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ by
$$m(U) = \begin{cases} |U| & \text{if } U \text{ is finite,} \\ \infty & \text{otherwise,} \end{cases}$$
for $U \in \mathcal{P}(S)$. Prove that $m$ is a measure.
48. Let $x, y, a_1, b_1, \ldots, a_n, b_n$ be $n$ real numbers such that $x \leq y$ and $a_i \leq b_i$ for $1 \leq i \leq n$. Prove, by induction on $n$, that if $[x, y] \subseteq \bigcup_{i=1}^{n} (a_i, b_i)$, then $y - x \leq \sum_{i=1}^{n} (b_i - a_i)$.
49. Let $(S, \mathcal{E}, m)$ be a measure space. Prove that if $\mathbf{s} = (S_0, S_1, \ldots)$ is a sequence of sets such that $\sum_i m(S_i) < \infty$, then $m(\liminf \mathbf{s}) = 0$ (*the Borel-Cantelli lemma*).

    **Solution:** Let $T_p = \bigcup_{i=p}^{\infty} S_i$ for $p \in \mathbb{N}$. By the subadditivity of $m$, we have $m(T_p) \leq \sum_{i=p}^{\infty} m(S_i)$, and therefore $\lim_{p \to \infty} m(T_p) = 0$ because of the convergence of the series $\sum_i m(S_i)$. Since $\liminf \mathbf{s} = \bigcap_{p=0}^{\infty} \bigcup_{i=p}^{\infty} S_i = \bigcap_{p=0}^{\infty} T_p$, it follows that $m(\liminf \mathbf{s}) \leq m(T_p)$ for every $p \in \mathbb{N}$, so $m(\liminf \mathbf{s}) \leq \inf_p m(T_p) = 0$, which implies $m(\liminf \mathbf{s}) = 0$.
50. Let $I$ be a bounded interval of $\mathbb{R}$. Prove that if $K$ is a compact subset of $\mathbb{R}$ such that $K \subseteq I$, then $\mu(I) = \mu(K) + \mu(I - K)$, where $\mu$ is the Lebesgue outer measure.
51. Let $\{(S_i, \mathcal{E}_i, m_i) \mid i \in I\}$ be a collection of measure spaces such that $S_i \mathcal{S}_| = \emptyset$ if $i \neq j$ for $i, j \in I$. Define the triplet $(\bigcup_{i \in I} S_i, \mathcal{E}, m)$, where
$$\mathcal{E} = \left\{ U \mid U \subseteq \bigcup_{i \in I} S_i, U \cap S_i \in \mathcal{E} \text{ for } i \in I \right\},$$
and $m : \mathcal{E} \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ is given by $m(U) = \sum_{i \in I} m_i(U \cap S_i)$ for $U \in \mathcal{E}$. Prove that $(\bigcup_{i \in I} S_i, \mathcal{E}, m)$ is a measure space and that $m(U)$ is finite if and only if there exists a countable subset $J$ of $I$ such that if $j \in J$, then $\mu_j$ is finite and $\mu_i = 0$ if $i \in I - J$.
52. The measure space $(S, \mathcal{E}, m)$ is *complete* if for every $W \in \mathcal{E}$ such that $m(W) = 0$, $U \subseteq W$ implies $U \in \mathcal{E}$. In Corollary 6.124 we saw that for every outer measure $\mu : \mathcal{P}(S) \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ the triple $(S, \mathcal{E}_\mu, \mu)$ is a measure space. Prove that this space is complete.
53. Let $(S, \mathcal{E}, m)$ be a measure space. Define $\mathcal{E}' = \{U \cup T \mid U \in \mathcal{E}, T \subseteq W \in \mathcal{E} \text{ and } m(W) = 0\}$, and $m' : \mathcal{E}' \longrightarrow \bar{\mathbb{R}}_{\geq 0}$ by $m'(U \cup T) = m(U)$ for every set $T$ such that $T \subseteq W \in \mathcal{E}$ and $m(W) = 0$.

a) Prove that $\mathcal{E}'$ is a $\sigma$-field that contains $\mathcal{E}$.
b) Prove that $m'$ is a measure. The measure $m'$ is known as the *completion of* $m$.

## Bibliographical Comments

There are several excellent classic references on general topology ([77, 40, 46, 44]). A very readable introduction to topology is [47]. Fundamental references on measure theory are [61, 100].

Pioneering work in applying topology in data mining has been done in [104, 83]. Important references in measure theory are [61, 100, 134].

# 7

# Frequent Item Sets and Association Rules

## 7.1 Introduction

Association rules have received lots of attention in data mining due to their many applications in marketing, advertising, inventory control, and many other areas.

A typical supermarket may well have several thousand items on its shelves. Clearly, the number of subsets of the set of items is immense. Even though a purchase by a customer involves a small subset of this set of items, the number of such subsets is very large. For example, even if we assume that no customer has more than five items in his shopping cart, there are $\sum_{i=1}^{5} \binom{10000}{i}$ possible contents of this cart, which corresponds to the subsets having no more than five items of a set that has 10,000 items, and this is indeed a large number!

The supermarket is interested in identifying associations between item sets; for example, it may be interested to know how many of the customers who bought bread and cheese also bought butter. This knowledge is important because if it turns out that many of the customers who bought bread and cheese also bought butter, the supermarket will place butter physically close to bread and cheese in order to stimulate the sales of butter. Of course, such a piece of knowledge is especially interesting when there is a substantial number of customers who buy all three items and a large fraction of those individuals who buy bread and cheese also buy butter.

We will formalize this problem and will explore its algorithmic aspects.

## 7.2 Frequent Item Sets

Suppose that $I$ is a finite set; we refer to the elements of $I$ as *items*.

**Definition 7.1.** *A* transaction data set on $I$ *is a function* $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$. *The set* $T(k)$ *is the* $k^{\text{th}}$ transaction of $T$. *The numbers* $1, \ldots, n$ *are the transaction identifiers (*tids*).*

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_7, © Springer-Verlag London Limited 2008

An example of a transaction set is the set of items present in the shopping cart of a consumer that completed a purchase in a store.

*Example 7.2.* The table below describes a transaction data set on the set of over-the-counter medicines in a drugstore.

| Trans. | Content |
|---|---|
| $T(1)$ | {Aspirin, Vitamin C} |
| $T(2)$ | {Aspirin, Sudafed} |
| $T(3)$ | {Tylenol} |
| $T(4)$ | {Aspirin, Vitamin C, Sudafed} |
| $T(5)$ | {Tylenol, Cepacol} |
| $T(6)$ | {Aspirin, Cepacol} |
| $T(7)$ | {Aspirin, Vitamin C} |

The same data set can be presented as a 0/1 table:

| | Aspirin | Vitamin C | Sudafed | Tylenol | Cepacol |
|---|---|---|---|---|---|
| $T(1)$ | 1 | 1 | 0 | 0 | 0 |
| $T(2)$ | 1 | 0 | 1 | 0 | 0 |
| $T(3)$ | 0 | 0 | 0 | 1 | 0 |
| $T(4)$ | 1 | 1 | 1 | 0 | 0 |
| $T(5)$ | 1 | 0 | 0 | 0 | 1 |
| $T(6)$ | 1 | 0 | 0 | 0 | 1 |
| $T(7)$ | 1 | 1 | 0 | 0 | 0 |

The entry in the row $T(k)$ and the column $i_j$ is set to 1 if $i_j \in T(k)$; otherwise, it is set to 0.

Example 7.2 shows that we have the option of two equivalent frameworks for studying frequent item sets: tables or transaction item sets.

Given a transaction data set $T$ on the set $I$, we would like to determine those subsets of $I$ that occur often enough as values of $T$.

**Definition 7.3.** *Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set on a set of items $I$. The* support count *of a subset $K$ of the set of items $I$ in $T$ is the number $suppcount_T(K)$ given by*

$$suppcount_T(K) = |\{k \mid 1 \leq k \leq n \text{ and } K \subseteq T(k)\}|.$$

*The support of an item set $K$ is the number*

$$supp_T(K) = \frac{suppcount_T(K)}{n}.$$

*Example 7.4.* For the transaction data set $T$ considered in Example 7.2, we have

$$suppcount_T(\{Aspirin, VitaminC\}) = 3$$

because $\{Aspirin, VitaminC\}$ is a subset of three of the sets $T(k)$. Therefore, $supp_T(\{Aspirin, VitaminC\}) = \frac{3}{7}$.

*Example 7.5.* Let $I = \{i_1, i_2, i_3, i_4\}$ be a collection of items. Consider the transaction data set $T$ given by

$$\begin{aligned} T(1) &= \{i_1, i_2\}, \\ T(2) &= \{i_1, i_3\}, \\ T(3) &= \{i_1, i_2, i_4\}, \\ T(4) &= \{i_1, i_3, i_4\}, \\ T(5) &= \{i_1, i_2\}, \\ T(6) &= \{i_3, i_4\}. \end{aligned}$$

Thus, the support count of the item set $\{i_1, i_2\}$ is 3; similarly, the support count of the item set $\{i_1, i_3\}$ is 2. Therefore, $supp_T(\{i_1, i_2\}) = \frac{1}{2}$ and $supp_T(\{i_1, i_3\}) = \frac{1}{3}$.

The following rather straightforward statement is fundamental for the study of frequent item sets.

**Theorem 7.6.** *Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set on a set of items $I$. If $K$ and $K'$ are two item sets, then $K' \subseteq K$ implies $supp_T(K') \geq supp_T(K)$.*

*Proof.* Note that every transaction that contains $K$ also contains $K'$. The statement follows immediately. □

If we seek those item sets that enjoy a minimum support level relative to a transaction data set $T$, then it is natural to start the process with the smallest nonempty item sets.

**Definition 7.7.** *An item set $K$ is* $\mu$-frequent *relative to the transaction data set $T$ if $supp_T(K) \geq \mu$.*

*We denote by $\mathcal{F}_T^\mu$ the collection of all $\mu$-frequent item sets relative to the transaction data set $T$ and by $\mathcal{F}_{T,r}^\mu$ the collection of $\mu$-frequent item sets that contain $r$ items for $r \geq 1$.*

Note that

$$\mathcal{F}_T^\mu = \bigcup_{r \geq 1} \mathcal{F}_{T,r}^\mu.$$

If $\mu$ and $T$ are clear from the context, then we may omit either or both adornments from this notation.

Let $I = \{i_1, \ldots, i_n\}$ be an item set that contains $n$ elements.

Denote by $\mathcal{G}_I = (\mathcal{P}(I), E)$ the Rymon tree of $\mathcal{P}(I)$. Recall that the root of the tree is $\emptyset$. A vertex $K = \{i_{p_1}, \ldots, i_{p_k}\}$ with $i_{p_1} < i_{p_2} < \cdots < i_{p_k}$ has $n - i_{p_k}$ children $K \cup \{j\}$, where $i_{p_k} < j \leq n$.

Let $\mathcal{S}_r$ be the collection of item sets that have $r$ elements. The next theorem suggests a technique for generating $\mathcal{S}_{r+1}$ starting from $\mathcal{S}_r$.

**Theorem 7.8.** *Let $\mathcal{G}$ be the Rymon tree of $\mathcal{P}(I)$, where $I = \{i_1, \ldots, i_n\}$. If $W \in \mathcal{S}_{r+1}$, where $r \geq 2$, then there exists a unique pair of distinct sets $U, V \in \mathcal{S}_r$ that has a common immediate ancestor $T \in \mathcal{S}_{r-1}$ in $\mathcal{G}$ such that $U \cap V \in \mathcal{S}_{r-1}$ and $W = U \cup V$.*

*Proof.* Let $u$ and $v$ be the two elements of $W$ that have the largest and the second-largest subscripts, respectively. Consider the sets $U = W - \{u\}$ and $V = W - \{v\}$. Both sets belong to $\mathcal{S}_r$. Moreover, $Z = U \cap V$ belongs to $\mathcal{S}_{r-1}$ because it consists of the first $r - 1$ elements of $W$. Note that both $U$ and $V$ are descendants of $Z$ and that $U \cup V = W$.

The pair $(U, V)$ is unique. Indeed, suppose that $W$ can be obtained in the same manner from another pair of distinct sets $U', V' \in \mathcal{S}_r$ such that $U'$ and $V'$ are immediate descendants of a set $Z' \in \mathcal{S}_{r-1}$. The definition of the Rymon tree $\mathcal{G}_I$ implies that $U' = Z' \cup \{i_m\}$ and $V' = Z' \cup \{i_q\}$, where the letters in $Z'$ are indexed by a number smaller than $\min\{m, q\}$. Then, $Z'$ consists of the first $r - 1$ symbols of $W$, so $Z' = Z$. If $m < q$, then $m$ is the second-highest index of a symbol in $W$ and $q$ is the highest index of a symbol in $W$, so $U' = U$ and $V' = V$. □

*Example 7.9.* Consider the Rymon tree of the collection $\mathcal{P}(\{i_1, i_2, i_3, i_4\})$ shown in Figure 7.1.

The set $\{i_1, i_3, i_4\}$ is the union of the sets $\{i_1, i_3\}$ and $\{i_1, i_4\}$ that have the common ancestor $\{i_1\}$.

Next we discuss an algorithm that allows us to compute the collection $\mathcal{F}_T^\mu$ of all $\mu$-frequent item sets for a transaction data set $T$. The algorithm is known as the *Apriori algorithm.*

We begin with the procedure `apriori_gen`, which starts with the collection $\mathcal{F}_{T,k}^\mu$ of frequent item sets for the transaction data set $T$ that contain $k$ elements and generates a collection $\mathcal{C}_{k+1}$ of sets of items that contains $\mathcal{F}_{T,k+1}^\mu$, the collection of the frequent item sets that have $k + 1$ elements. The justification for this procedure is based on the next statement.

**Theorem 7.10.** *Let $T$ be a transaction data set on a set of items $I$ and let $k \in \mathbb{N}$ such that $k > 1$.*

*If $W$ is a $\mu$-frequent item set and $|W| = k + 1$, then there exists a $\mu$-frequent item set $Z$ and two items $i_m$ and $i_q$ such that $|Z| = k - 1$, $Z \subseteq W$, $W = Z \cup \{i_m, i_q\}$, and both $Z \cup \{i_m\}$ and $Z \cup \{i_q\}$ are $\mu$-frequent item sets.*

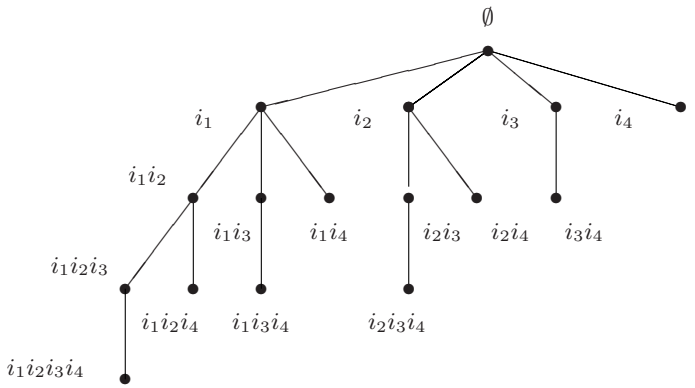



**Fig. 7.1.** Rymon tree for $\mathcal{P}(\{i_1, i_2, i_3, i_4\})$.

*Proof.* If $W$ is an item set such that $|W| = k+1$, then we already know that $W$ is the union of two subsets $U$ and $V$ of $I$ such that $|U| = |V| = k$ and that $Z = U \cap V$ has $k-1$ elements. Since $W$ is a $\mu$-frequent item set and $Z, U, V$ are subsets of $W$, it follows that each of these sets is also a $\mu$-frequent item set. □

Note that the reciprocal statement of Theorem 7.10 is not true, as the next example shows.

*Example 7.11.* Let $T$ be the transaction data set introduced in Example 7.5. Note that both $\{i_1, i_2\}$ and $\{i_1, i_3\}$ are $\frac{1}{3}$-frequent item sets; however,

$$supp_T(\{i_1, i_2, i_3\}) = 0,$$

so $\{i_1, i_2, i_3\}$ fails to be a $\frac{1}{3}$-frequent item set.

The procedure `apriori_gen` mentioned above is Algorithm 7.12. This procedure starts with the collection of item sets $\mathcal{F}_{T,k}$ and produces a collection of item sets $\mathcal{C}_{T,k+1}$ that includes the collection of item sets $\mathcal{F}_{T,k+1}$ of frequent item sets having $k+1$ elements.

**Algorithm 7.12 (Procedure `apriori_gen`)**
**Input:** a minimum support $\mu$, the collection $\mathcal{F}^{\mu}_{T,k}$ of frequent item sets having $k$ elements;
**Output:** the set of candidate frequent item sets $\mathcal{C}^{\mu}_{T,k+1}$;
**Method:**

set $j = 1$;
$\mathcal{C}^{\mu}_{T,j+1} = \emptyset$;
for each $L, M \in \mathcal{F}^{\mu}_{T,k}$ such that
    $L \neq M$ and $L \cap M \in \mathcal{F}^{\mu}_{T,k-1}$ do
    add $L \cup M$ to $\mathcal{C}^{\mu}_{T,k+1}$;
remove all sets $K$ in $\mathcal{C}^{\mu}_{T,k+1}$ where
    there is a subset of $K$ containing $k$ elements
    that does not belong to $\mathcal{F}^{\mu}_{T,k}$.

Note that in `apriori_gen` no access to the transaction data set is needed.

The *Apriori* algorithm 7.13 operates on "levels." Each level $k$ consists of a collection $\mathcal{C}^{\mu}_{T,k}$ of candidate item sets of $\mu$-frequent item sets. To build the initial collection of candidate item sets $\mathcal{C}^{\mu}_{T,1}$, every single item set is considered for membership in $\mathcal{C}^{\mu}_{T,1}$. The initial set of frequent item sets consists of those singletons that pass the minimal support test. The algorithm alternates between a candidate generation phase (accomplished by using `apriori_gen`) and an evaluation phase that involves a data set scan and is therefore the most expensive component of the algorithm.

**Algorithm 7.13 (The Apriori Algorithm)**
**Input:** transaction data set $T$ and a minimum support $\mu$;
**Output:** the collection $\mathcal{F}^{\mu}_{T}$ of $\mu$-frequent item sets;
**Method:** $\mathcal{C}^{\mu}_{T,1} = \{\{i\} \mid i \in I\}$;
    set $i = 1$;
    **while** $(\mathcal{C}^{\mu}_{T,i} \neq \emptyset)$ **do**
        /* evaluation phase */
        $\mathcal{F}^{\mu}_{T,i} = \{L \in \mathcal{C}^{\mu}_{T,i} \mid supp_T(L) \geq \mu\}$;
        /* candidate generation */
        $\mathcal{C}^{\mu}_{T,i+1} =$ `apriori_gen`$(\mathcal{F}^{\mu}_{T,i})$;
        $i++$;
    **end while**;
    **output** $\mathcal{F}^{\mu}_{T} = \bigcup_{j<i} \mathcal{F}^{\mu}_{T,j}$

*Example 7.14.* Let $T$ be the data set given by

| | $i_1$ | $i_2$ | $i_3$ | $i_4$ | $i_5$ |
|---|---|---|---|---|---|
| $T(1)$ | 1 | 1 | 0 | 0 | 0 |
| $T(2)$ | 0 | 1 | 1 | 0 | 0 |
| $T(3)$ | 1 | 0 | 0 | 0 | 1 |
| $T(4)$ | 1 | 0 | 0 | 0 | 1 |
| $T(5)$ | 0 | 1 | 1 | 0 | 1 |
| $T(6)$ | 1 | 1 | 1 | 1 | 1 |
| $T(7)$ | 1 | 1 | 1 | 0 | 0 |
| $T(8)$ | 0 | 1 | 1 | 1 | 1 |

The support counts of various subsets of $I = \{i_1, \ldots, i_5\}$ are given below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $i_1$ | | $i_2$ | | $i_3$ | | $i_4$ | | $i_5$ | |
| 5 | | 6 | | 5 | | 2 | | 5 | |
| $i_1i_2$ | $i_1i_3$ | $i_1i_4$ | $i_1i_5$ | $i_2i_3$ | $i_2i_4$ | $i_2i_5$ | $i_3i_4$ | $i_3i_5$ | $i_4i_5$ |
| 3 | 2 | 1 | 3 | 5 | 2 | 3 | 2 | 3 | 2 |
| $i_1i_2i_3$ | $i_1i_2i_4$ | $i_1i_2i_5$ | $i_1i_3i_4$ | $i_1i_3i_5$ | $i_1i_4i_5$ | $i_2i_3i_4$ | $i_2i_3i_5$ | $i_2i_4i_5$ | $i_3i_4i_5$ |
| 2 | 1 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 2 |
| $i_1i_2i_3i_4$ | | $i_1i_2i_3i_5$ | | $i_1i_2i_4i_5$ | | $i_1i_3i_4i_5$ | | $i_2i_3i_4i_5$ | |
| 1 | | 1 | | 1 | | 1 | | 2 | |
| | | | | $i_1i_2i_3i_4i_5$ | | | | | |
| | | | | 0 | | | | | |

Starting with $\mu = 0.25$ and $\mathcal{F}^{\mu}_{T,0} = \{\emptyset\}$, the Apriori algorithm computes the following sequence of sets:

$$
\begin{aligned}
\mathcal{C}^{\mu}_{T,1} &= \{i_1, i_2, i_3, i_4, i_5\},\\
\mathcal{F}^{\mu}_{T,1} &= \{i_1, i_2, i_3, i_4, i_5\},\\
\mathcal{C}^{\mu}_{T,2} &= \{i_1i_2, i_1i_3, i_1i_4, i_1i_5, i_2i_3, i_2i_4, i_2i_5, i_3i_4, i_3i_5, i_4i_5\},\\
\mathcal{F}^{\mu}_{T,2} &= \{i_1i_2, i_1i_3, i_1i_5, i_2i_3, i_2i_4, i_2i_5, i_3i_4, i_3i_5, i_4i_5\},\\
\mathcal{C}^{\mu}_{T,3} &= \{i_1i_2i_3, i_1i_2i_5, i_1i_3i_5, i_2i_3i_4, i_2i_3i_5, i_2i_4i_5, i_3i_4i_5\},\\
\mathcal{F}^{\mu}_{T,3} &= \{i_1i_2i_3, i_2i_3i_4, i_2i_3i_5, i_2i_4i_5, i_3i_4i_5\},\\
\mathcal{C}^{\mu}_{T,4} &= \{i_2i_3i_4i_5\},\\
\mathcal{F}^{\mu}_{T,4} &= \{i_2i_3i_4i_5\},\\
\mathcal{C}^{\mu}_{T,5} &= \emptyset.
\end{aligned}
$$

Thus, the algorithm will output the collection

$$
\begin{aligned}
\mathcal{F}^{\mu}_{T} &= \bigcup_{i=1}^{4} \mathcal{F}^{\mu}_{T,i}\\
&= \{i_1, i_2, i_3, i_4, i_5, i_1i_2, i_1i_3, i_1i_5, i_2i_3, i_2i_4, i_2i_5, i_3i_4, i_3i_5, i_4i_5,\\
&\quad i_1i_2i_3, i_2i_3i_4, i_2i_3i_5, i_2i_4i_5, i_3i_4i_5, i_2i_3i_4i_5\}.
\end{aligned}
$$

## 7.3 Borders of Collections of Sets

Let $\mathcal{I}$ be a collection of sets such that $\mathcal{I} \subseteq \mathcal{P}(I)$, where $I$ is a set.

**Definition 7.15.** *The* border of $\mathcal{I}$ *is the collection*

$$BD(\mathcal{I}) = \{L \in \mathcal{P}(I) \mid U \subset L \text{ implies } U \in \mathcal{I} \text{ and } L \subset V \text{ implies } V \notin \mathcal{I}\}.$$

*The* positive border of $\mathcal{I}$ *is the collection*

$$BD^+(\mathcal{I}) = BD(\mathcal{I}) \cap \mathcal{I}$$
$$= \{L \in \mathcal{I} \mid U \subset L \textit{ implies } U \in \mathcal{I} \textit{ and } L \subset V \textit{ implies } V \notin \mathcal{I}\},\}$$

*while the negative border is*

$$BD^-(\mathcal{I}) = BD(\mathcal{I}) - \mathcal{I}$$
$$= \{L \in \mathcal{P}(I) - \mathcal{I} \mid U \subset L \textit{ implies } U \in \mathcal{I} \textit{ and } L \subset V \textit{ implies } V \notin \mathcal{I}\}.$$

*Clearly, we have* $BD(\mathcal{I}) = BD^+(\mathcal{I}) \cup BD^-(\mathcal{I})$.

If $\mathcal{I}$ is a hereditary collection of sets (see Definition 1.14), then the positive and the negative borders of $\mathcal{I}$ are given by

$$BD^+(\mathcal{I}) = \{L \in \mathcal{I} \mid L \subset V \text{ implies } V \notin \mathcal{I}\}$$

and

$$BD^-(\mathcal{I}) = \{L \in \mathcal{P}(I) - \mathcal{I} \mid U \subset L \text{ implies } U \in \mathcal{I}\},$$

respectively. Thus, for a hereditary collection of subsets $\mathcal{I}$, the positive border consists of the maximal subsets of $\mathcal{I}$, while the negative border of $\mathcal{I}$ consists of the minimal subsets of the collection $\mathcal{P}(I) - \mathcal{I}$.

Note that if $\mathcal{I}$ and $\mathcal{I}'$ are two hereditary collections of subsets of $I$ and $BD^+(\mathcal{I}) = BD^+(\mathcal{I}')$, then $\mathcal{I} = \mathcal{I}'$. Indeed, if $K \in \mathcal{I}$, one of the following two cases may occur:

1. If $K$ is not a maximal set of $\mathcal{I}$, then there is a maximal set $H$ of $\mathcal{I}$ such that $K \subset H$. Since $H \in BD^+(\mathcal{I}) = BD^+(\mathcal{I}')$, it follows that $H \in \mathcal{I}'$, hence $K \in \mathcal{I}'$ because $\mathcal{I}'$ is hereditary.
2. If $K$ is a maximal set of $\mathcal{I}$, then $K \in BD^+(\mathcal{I}) = BD^+(\mathcal{I}')$; hence, $K \in \mathcal{I}'$.

In either case $K \in \mathcal{I}'$, so $\mathcal{I} \subseteq \mathcal{I}'$. The reverse inclusion can be proven in a similar way, so $\mathcal{I} = \mathcal{I}'$.

Similarly, we can show that for two hereditary collections $\mathcal{I}, \mathcal{I}'$ of subsets of $I$, $BD^-(\mathcal{I}) = BD^-(\mathcal{I}')$ implies $\mathcal{I} = \mathcal{I}'$. Indeed, suppose that $K \in \mathcal{I} - \mathcal{I}'$. Since $K \notin \mathcal{I}'$, there exists a minimal subset $V$ of $K$ such that $V \notin \mathcal{I}'$ and each of its subsets is in $\mathcal{I}'$. The set $V$ belongs to the negative border $BD^-(\mathcal{I}')$ and, therefore to $BD^-(\mathcal{I})$. This leads to a contradiction because $K \in \mathcal{I}$ and $V$ is subset of $K$ does not belong to $\mathcal{I}$, thereby contradicting the fact that $\mathcal{I}$ is a hereditary family of sets.

Since no such set $K$ may exist, it follows that $\mathcal{I} \subseteq \mathcal{I}'$. The reverse inclusion can be be shown in the same manner.

Borders of collections of sets play an important role in the study of the Apriori algorithm. Observe, for example, that after computing the collection $\mathcal{F}^{\mu}_{T,3} = \{i_1i_2i_3, i_2i_3i_4, i_2i_3i_5, i_2i_4i_5, i_3i_4i_5\}$ in Example 7.14, the candidate set $\mathcal{C}^{\mu}_{T,4} = \{i_2i_3i_4i_5\}$ is the negative border of $\mathcal{F}^{\mu}_{T,3}$. In general, $\mathcal{C}^{\mu}_{T,i+1}$ is the negative border $BD^-(\mathcal{F}^{\mu}_{T,i})$.

For the same example, the negative and the positive borders of the collection of frequent sets $\mathcal{F}^{\mu}_{T}$ are given by

$$BD^+(\mathcal{F}_T^\mu) = \{i_1i_5, i_1i_2i_3, i_2i_3i_4i_5\},$$
$$BD^-(\mathcal{F}_T^\mu) = \{i_1i_4, i_1i_2i_5, i_1i_3i_5\},$$

respectively. Clearly, $BD^+(\mathcal{F}_T^\mu)$ consists of the maximal $\mu$-frequent item sets, while $BD^-(\mathcal{F}_T^\mu)$ consists of the minimal $\mu$-infrequent item sets.

The time complexity of the Apriori algorithm is dominated by the number of accesses to the data set $T$ that is required for computing the support of candidate item sets.

**Theorem 7.16.** *The Apriori algorithm performs* $|\mathcal{F}_T^\mu| + |BD^-(\mathcal{F}_T^\mu)|$ *support computations.*

*Proof.* The Apriori algorithm selects the $\mu$-frequent item sets from among the candidate item sets, and for each candidate item set it must perform a support computation. A number of $|\mathcal{F}_T^\mu|$ candidate sets turn out to be $\mu$-frequent, so the algorithm will perform $|\mathcal{F}_T^\mu|$ computations for these sets. On the other hand, a candidate set $C$ is not retained as a $\mu$-frequent set if and only if all its subsets are $\mu$-frequent (a requirement of `apriori-gen`) and $C$ itself is not $\mu$-frequent, which means that none of its supersets are $\mu$-frequent. This happens if and only if $C$ belongs to the negative border of $\mathcal{F}_T^\mu$. Thus, the total number of support computations is $|\mathcal{F}_T^\mu| + |BD^-(\mathcal{F}_T^\mu)|$. □

**Theorem 7.17.** *Let $I$ be a set and let $\mathcal{I}$ be a hereditary family of subsets of $I$. Consider the collection of sets*

$$\mathcal{E} = \{I - L \mid L \in BD^+(\mathcal{I})\}$$

*and the hypergraph $\mathcal{H} = (I, \mathcal{E})$. Then, the collection of minimal transversals of the hypergraph $\mathcal{E}$ equals $BD^-(\mathcal{I})$, the negative border of $\mathcal{I}$.*

*Proof.* The following statements concerning a subset $X$ of $I$ are easily seen to be equivalent:

(i) $X$ is a transversal of $\mathcal{H}$.
(ii) $X \cap Y \neq \emptyset$ for every $Y \in \mathcal{E}$.
(iii) $X \cap (I - L) \neq \emptyset$ for every $L \in BD^+(\mathcal{I})$.
(iv) $X$ is not included in any maximal set $L$ of $\mathcal{I}$.
(v) $X \notin \mathcal{I}$.

Thus, $X$ is a transversal of $\mathcal{H}$ if and only if $X \notin \mathcal{I}$. Consequently, $X$ is a minimal transversal of $\mathcal{H}$ if and only if $X$ is a minimal set with the property that $X \notin \mathcal{I}$, which means that $X \in BD^-(\mathcal{I})$. □

## 7.4 Association Rules

**Definition 7.18.** *An* association rule *on an item set $I$ is a pair of nonempty disjoint item sets $(X, Y)$.*

Note that if $|I| = n$, then there exist $3^n - 2^{n+1} + 1$ association rules on $I$. Indeed, suppose that the set $X$ contains $k$ elements; there are $\binom{n}{k}$ ways of choosing $X$. Once $X$ is chosen, $Y$ can be chosen among the remaining $2^{n-k}-1$ nonempty subsets of $I - X$. In other words, the number of association rules is

$$\sum_{k=1}^{n} \binom{n}{k}(2^{n-k} - 1) = \sum_{k=1}^{n} \binom{n}{k}2^{n-k} - \sum_{k=1}^{n} \binom{n}{k}.$$

By taking $x = 2$ in the equality

$$(1 + x)^n = \sum_{k=0}^{n} \binom{n}{k}x^{n-k},$$

we obtain

$$\sum_{k=1}^{n} \binom{n}{k}2^{n-k} = 3^n - 2^n.$$

Since $\sum_{k=1}^{n} \binom{n}{k} = 2^n - 1$ the desired equality follows immediately. The number of association rules can be quite considerable even for small values of $n$. For example, for $n = 10$, we have $3^{10} - 2^{11} + 1 = 57002$ association rules.

An association rule $(X, Y)$ is denoted by $X \Rightarrow Y$. The confidence of $X \Rightarrow Y$ is the number

$$conf_T(X \Rightarrow Y) = \frac{supp_T(XY)}{supp_T(X)}.$$

**Definition 7.19.** *An association rule holds in a transaction data set $T$ with support $\mu$ and confidence $c$ if $supp_T(XY) \geq \mu$ and $conf_T(X \Rightarrow Y) \geq c$.*

Once a $\mu$-frequent item set $Z$ is identified, we need to examine the support levels of the subsets $X$ of $Z$ to ensure that an association rule of the form $X \Rightarrow Z - X$ has a sufficient level of confidence, $conf_T(X \Rightarrow Z - X) = \frac{\mu}{supp_T(X)}$. Observe that $supp_T(X) \geq \mu$ because $X$ is a subset of $Z$. To obtain a high level of confidence for $X \Rightarrow Z - X$, the support of $X$ must be as small as possible.

Clearly, if $X \Rightarrow Z - X$ does not meet the level of confidence, then it is pointless to look for rules of the form $X' \Rightarrow Z - X'$ among the subsets $X'$ of $X$.

*Example 7.20.* Let $T$ be the transaction data set introduced in Example 7.14. We saw that the item set $L = i_2i_3i_4i_5$ has support count equal to 2 and therefore $supp_T(L) = 0.25$. This allows us to obtain the following association rules having three item sets in their antecedent that are subsets of $L$:

| Rule | $suppcount_T(X)$ | $conf_T(X \Rightarrow Y)$ |
|---|---|---|
| $i_2i_3i_4 \Rightarrow i_5$ | 2 | 1 |
| $i_2i_3i_5 \Rightarrow i_4$ | 3 | $\frac{2}{3}$ |
| $i_2i_4i_5 \Rightarrow i_3$ | 2 | 1 |
| $i_3i_4i_5 \Rightarrow i_2$ | 2 | 1 |

Note that $i_2i_3i_4 \Rightarrow i_5$, $i_2i_4i_5 \Rightarrow i_3$, and $i_3i_4i_5 \Rightarrow i_2$ have 100% confidence. We refer to such rules as *exact association rules*.

The rule $i_2i_3i_5 \Rightarrow i_4$ has confidence $\frac{2}{3}$. It is clear that the confidence of rules of the form $U \Rightarrow V$ with $U \subseteq i_2i_3i_5$ and $UV = L$ will be lower than $\frac{2}{3}$ since $supp_T(U)$ is at least 3. Indeed, the possible rules of this form are:

| Rule | $suppcount_T(X)$ | $conf_T(X \Rightarrow Y)$ |
|---|---|---|
| $i_2i_3 \Rightarrow i_4i_5$ | 5 | $\frac{2}{5}$ |
| $i_2i_5 \Rightarrow i_3i_4$ | 3 | $\frac{2}{3}$ |
| $i_3i_5 \Rightarrow i_2i_4$ | 3 | $\frac{2}{3}$ |
| $i_2 \Rightarrow i_3i_4i_5$ | 6 | $\frac{1}{3}$ |
| $i_3 \Rightarrow i_2i_4i_5$ | 5 | $\frac{1}{5}$ |
| $i_5 \Rightarrow i_2i_3i_4$ | 5 | $\frac{2}{5}$ |

Obviously, if we seek association rules having a confidence larger than $\frac{2}{3}$, no such rule $U \Rightarrow V$ can be found such that $U$ is a subset of $i_2i_3i_5$.

Suppose, for example, that we seek association rules $U \Rightarrow V$ that have a minimal confidence of 80%. We need to examine subsets $U$ of the other sets, $i_2i_3i_4$, $i_2i_4i_5$, or $i_3i_4i_5$, which are not subsets of $i_2i_3i_5$ (since the subsets of $i_2i_3i_5$ cannot yield levels of confidence higher than $\frac{2}{3}$). There are five such sets:

| Rule | $suppcount_T(X)$ | $conf_T(X \Rightarrow Y)$ |
|---|---|---|
| $i_2i_4 \Rightarrow i_3i_5$ | 2 | 1 |
| $i_3i_4 \Rightarrow i_2i_5$ | 2 | 1 |
| $i_4i_5 \Rightarrow i_2i_3$ | 2 | 1 |
| $i_3i_4 \Rightarrow i_2i_5$ | 2 | 1 |
| $i_4 \Rightarrow i_2i_3i_5$ | 2 | 1 |

Indeed, all these sets yield exact rules, that is, rules having 100% confidence.

Many transaction data sets produce a huge number of frequent item sets and therefore a huge number of association rules, particularly when the levels of support and confidence required are relatively low. Moreover, it is well-known (see [132]) that limiting the analysis of association rules to the support/confidence framework can lead to dubious conclusions. The data mining literature contains many references that attempt to derive interestingness measures for association rules in order to focus data analysis of those rules that may be more relevant (see [107, 4, 7, 23, 74, 68]).

## 7.5 Levelwise Algorithms and Posets

This section focuses on the levelwise algorithms, a powerful and elegant generalization of the Apriori algorithm that was introduced in [95].

Let $(P, \leq)$ be a partially ordered set and let $Q$ be a subset of $P$.

**Definition 7.21.** *The border of $Q$ is the set*

$$BD(Q) = \{p \in P \mid u < p \text{ implies } u \in Q \text{ and } p < v \text{ implies } v \notin Q\}.$$

*The positive border of $Q$ is the set:*

$$BD^+(Q) = BD(Q) \cap Q,$$

*while the negative border of $Q$ is*

$$BD^-(Q) = BD(Q) - Q.$$

Clearly, we have $BD(Q) = BD^+(Q) \cup BD^-(Q)$.

An alternative terminology exists that makes use of the terms *generalization* and *specialization*. If $r, p \in P$ and $r < p$, then we say that $r$ is a *generalization* of $p$ or that $p$ is a *specialization* of $r$. Thus, the border of a set $Q$ consists of those elements $p$ of $P$ such that all of their generalizations are in $Q$ and none of their specializations is in $Q$.

**Theorem 7.22.** *Let $(P, \leq)$ be a partially ordered set. If $Q$ and $Q'$ are two disjoint subsets of $P$, then $BD(Q \cup Q') \subseteq BD(Q) \cup BD(Q')$.*

*Proof.* Let $p \in BD(Q \cup Q')$. Suppose that $u < p$, so $u \in Q \cup Q'$. Since $Q$ and $Q'$ are disjoint, we have either $u \in Q$ or $u \in Q'$. On the other hand, if $p < v$, then $v \notin Q \cup Q'$, so $v \notin Q$ and $v \notin Q'$. Thus, we have $p \in BD(Q) \cup BD(Q')$. □

The notion of a hereditary subset of a poset is an immediate generalization of the notion of a hereditary family of sets.

**Definition 7.23.** *A subset $Q$ of a poset $(P, \leq)$ is said to be hereditary if $p \in Q$ and $r \leq p$ imply $r \in Q$.*

**Theorem 7.24.** *If $Q$ is a hereditary subset of a poset $(P, \leq)$, then the positive and the negative borders of $Q$ are given by*

$$BD^+(Q) = \{p \in Q \mid p < v \text{ implies } v \notin Q\}$$

*and*

$$BD^-(Q) = \{p \in P - Q \mid u < p \text{ implies } u \in Q\},$$

*respectively.*

*Proof.* Let $t$ be an element of the positive border $BD^+(Q) = BD(Q) \cap Q$. We have $t \in Q$ and $t < v$ implies $v \notin Q$ because $t \in BD(Q)$.

Conversely, suppose that $t$ is an element of $Q$ such that $t < v$ implies $v \notin Q$. Since $Q$ is hereditary, $u < t$ implies $u \in Q$, so $t \in BD(Q)$. Therefore, $t \in BD(Q) \cap Q = BD^+(Q)$.

Now let $s$ be an element of the negative border of $Q$; that is, $s \in BD(Q) - Q$. We have immediately $s \in P - Q$. If $u < s$, then $u \in Q$, because $Q$ is hereditary. Thus, $BD^-(Q) \subseteq \{p \in P - Q \mid u < p \text{ implies } u \in Q\}$.

Conversely, suppose that $s \in P - Q$ and $u < s$ implies $u \in Q$. If $s < v$, then $v$ cannot belong to $Q$ because this would entail $s \in Q$ due to the hereditary property of $Q$. Consequently, $s \in BD(Q)$ and so $s \in BD(Q) - Q = BD^-(Q)$. □

Theorem 7.24 can be paraphrased by saying that for a hereditary subset $Q$ of $P$ the positive border consists of the maximal elements of $Q$, while the negative border of $Q$ consists of the minimal elements of $P - Q$.

Note that if $Q$ and $Q'$ are two hereditary subsets of $P$ and $BD^+(Q) = BD^+(Q')$, then $Q = Q'$. Indeed, if $z \in P$, one of the following two cases may occur:

1. If $z$ is not a maximal element of $Q$, then there is a maximal element $w$ of $Q$ such that $z < w$. Since $w \in BD^+(Q) = BD^+(Q')$, it follows that $w \in Q'$; hence $z \in Q'$ because $Q'$ is hereditary.
2. If $z$ is a maximal element of $Q$, then $z \in BD^+(Q) = BD^+(Q')$, hence $z \in Q'$.

In either case $z \in Q'$, so $Q \subseteq Q'$. The reverse inclusion can be proven in a similar way, so $Q = Q'$.

Similarly, we can show that for two hereditary collections $Q$ and $Q'$ of subsets of $I$, $BD^-(Q) = BD^-(Q')$ implies $Q = Q'$. Indeed, suppose that $z \in Q - Q'$. Since $z \notin Q'$, there exists a minimal element $v$ such that $v \notin Q'$ and each of its lower bounds is in $Q'$. Since $v$ belongs to the negative border $BD^-(Q')$, it follows that $v \in BD^-(Q)$. This leads to a contradiction because $z \in Q$ and $v$ (for which we have $v < z$) does not, thereby contradicting the fact that $Q$ is a hereditary subset. Since no such $z$ may exist, it follows that $Q \subseteq Q'$. The reverse inclusion can be shown in the same manner.

**Definition 7.25.** *Let $\mathcal{D}$ be a relational database, $\mathcal{S}_\mathcal{D}$ be the set of states of $\mathcal{D}$, and $(B, \leq, h)$ be a ranked poset referred to as the ranked poset of objects.*

*A query is a function $q : \mathcal{S}_\mathcal{D} \times B \longrightarrow \{0, 1\}$ such that $D \in \mathcal{S}_\mathcal{D}$, $b \leq b'$, and $q(D, b') = 1$ imply $q(D, b) = 1$.*

Definition 7.25 is meant to capture the framework of the Apriori algorithm for identification of frequent item sets. As was shown in [95], this framework can capture many other situations.

*Example 7.26.* Let $\mathcal{D}$ be a database that contains a tabular variable $(T, H)$ and let $\theta = (T, H, \rho)$ be the table that is the current value of $(T, H)$ contained by the current state $D$ of $\mathcal{D}$.

The graded poset $(B, \leq, h)$ is $(\mathcal{P}(H), \subseteq, h)$, where $h(X) = |X|$. Given a number $\mu$, the query is defined by

$$q(D, K) = \begin{cases} 1 & \text{if } supp_T(K) \leq \mu, \\ 0 & \text{otherwise.} \end{cases}$$

Since $K \subseteq K'$ implies $supp_T(K') \leq supp_T(K)$, it follows that $q$ satisfies the condition of Definition 7.25.

*Example 7.27.* As in Example 7.26, let $\mathcal{D}$ be a database that contains a tabular variable $(T, H)$, and let $\theta = (T, H, \rho)$ be the table that is the current value of $(T, H)$ contained by the current state $D$ of $\mathcal{D}$. The graded poset $(\mathcal{P}(H), \supseteq, g)$ is the dual of the graded poset considered in Example 7.26, where $g(K) = |H| - |K|$. If $L$ is a set of attributes, the function $q_L$ is defined by

$$q_L(D, K) = \begin{cases} 1 & \text{if } K \to L \text{ holds in } \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Note that if $K' \subseteq K$ and $D$ satisfies the functional dependency $K' \to L$, then $D$ satisfies $K \to L$. Thus, $q$ is a query in the sense of Definition 7.25.

**Definition 7.28.** *The set of interesting objects for the state $D$ of the database and the query $q$ is given by*

$$INT(D, q) = \{b \in B \mid q(D, b) = 1\}.$$

Note that the set of interesting objects is a hereditary set $(B, \leq)$. Indeed, if $b \in INT(D, q)$ and $c \leq b$, then $c \in INT(D, q)$ according to Definition 7.25. Thus,

$$BD^+(INT(D, q)) = \{b \in INT(D, q) \mid b < v \text{ implies } v \notin INT(D, q)\},$$
$$BD^-(INT(D, q)) = \{b \in B - INT(D, q) \mid u < b \text{ implies } u \in INT(D, q)\}.$$

In other words, $BD^+(INT(D, q))$ is the set of maximal objects that are interesting, while $BD^-(INT(D, q))$ is the set of minimal objects that are not interesting.

Algorithm 7.29, which we discuss next, is a general algorithm that seeks to compute the set of interesting objects for a state of a database. The algorithm is known as the *levelwise algorithm* because it identifies these objects by successively scanning the levels of the graded poset of objects.

If $L_0, L_1, \ldots$ are the levels of the graded poset $(B, \leq, h)$, then the algorithm begins by examining all objects located on the initial level. The set of interesting objects located on the level $L_i$ is denoted by $\mathcal{F}_i$; for each level $L_i$, the computation of $\mathcal{F}_i$ is preceded by a computation of the set of potentially interesting objects $\mathcal{C}_i$ referred to as the set of *candidate objects*.

The first set of candidate objects, $\mathcal{C}_1$, coincides with the level $L_i$. Only the interesting objects on this level are retained for the set $\mathcal{F}_1$.

The next set of candidate objects, $\mathcal{C}_{i+1}$, is constructed by examining the level $L_{i+1}$ and keeping those objects $b$ having all their subobjects $c$ in the interesting sets of the previous levels.

**Algorithm 7.29 (General Levelwise Algorithm)**
**Input:** a database state $D$, a graded poset of objects $(B, \leq, h)$, and a query $q$;
**Output:** the set of interesting objects for $D$;
**Method:** $\mathcal{C}_1 = L_1$;
$i = 1$;
**while** $(\mathcal{C}_i \neq \emptyset)$ **do**
/* evaluation phase */
$\mathcal{F}_i = \{b \in \mathcal{C}_i \mid q(D, b) = 1\}$;
/* candidate generation */
$\mathcal{C}_{i+1} = \{b \in L_{i+1} \mid c < b \text{ implies } c \in \bigcup_{j \leq i} \mathcal{F}_j\} - \bigcup_{j \leq i} \mathcal{C}_j$
$i++$;
**end while**;
output $\bigcup_{j<i} \mathcal{F}_j$

*Example 7.30.* For frequent item sets, we can work in the framework described in Example 7.26. The algorithm, which is essentially the Apriori algorithm described in Section 7.2, goes through the **while** loop no more than $k + 1$ times, where

$$k = \max\{|X| \mid X \subseteq H, supp_T(X) > \mu\}.$$

*Example 7.31.* In Example 7.27, we defined the grading query $q_L$ as

$$q_L(D, K) = \begin{cases} 1 & \text{if } K \to L \text{ holds in } \theta, \\ 0 & \text{otherwise,} \end{cases}$$

for $K \in \mathcal{P}(H)$. The levelwise algorithm allows us to identify those subsets $K$ such that a table $\theta = (T, H, \rho)$ satisfies the functional dependency $K \to L$. The first level consists of all subsets $K$ of $H$ that have $|H| - 1$ attributes. There are, of course, $|H| - 1$ such subsets, and the set $\mathcal{F}_1$ will contain all these sets such that $K \to H$ is satisfied. Successive levels contain sets that have fewer and fewer attributes. Level $L_i$ contains sets that have $|H| - i$ attributes.

The algorithm will go through the **while** loop at most $1 + |H - K|$ times, where $K$ is the smallest set such that $K \to L$ holds.

Observe that the computation of $\mathcal{C}_{i+1}$ in the generic levelwise algorithm

$$\mathcal{C}_{i+1} = \left\{ b \in L_{i+1} \mid c < b \text{ implies } c \in \bigcup_{j \leq i} \mathcal{F}_j \right\} - \bigcup_{j \leq i} \mathcal{C}_j$$

can be written as

$$\mathcal{C}_{i+1} = BD^-\left(\bigcup_{j \leq i} \mathcal{F}_j\right) - \bigcup_{j \leq i} \mathcal{C}_j.$$

This shows that the set of candidate objects at level $L_{i+1}$ is the negative border of the interesting sets located on the lower level, excluding those objects that have already been evaluated.

The most expensive component of the levelwise algorithm is the evaluation of $q(D, b)$ since this requires a scan of the database state $D$. Clearly, we need to evaluate this function for each candidate element, so we will require $|\bigcup_{i=1}^{\ell} \mathcal{C}_i|$ evaluations, where $\ell$ is the number of levels that are scanned. Some of these evaluations will result in including the evaluated object $b$ in the set $\mathcal{F}_i$. Objects that will not be included in $INT(D, q)$ are such that any of their generalizations are in $INT(D, q)$, even though they fail to belong to this set. They belong to $BD^-(INT(D, q))$. Thus, the levelwise algorithm performs $|INT(D, q)| + |BD^-(INT(D, q))|$ evaluations of $q(D, b)$.

## 7.6 Lattices and Frequent Item Sets

Galois connections discussed in Section 5.4 are useful in the study of frequent item sets. This approach was introduced for the first time in [105].

Let $I$ be a set of items and $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set. Denote by $D$ the set of transaction identifiers $D = \{1, \ldots, n\}$. The functions $items_T : \mathcal{P}(D) \longrightarrow \mathcal{P}(I)$ and $tids_T : \mathcal{P}(I) \longrightarrow \mathcal{P}(D)$ are defined by

$$items_T(E) = \bigcap \{T(k) \mid k \in E\},$$
$$tids_T(H) = \{k \in D \mid H \subseteq T(k)\},$$

for every $E \in \mathcal{P}(D)$ and every $H \in \mathcal{P}(I)$.

Note that $suppcount_T(H) = |tids_T(H)|$ for every $H \in \mathcal{P}(I)$.

**Theorem 7.32.** *Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set. The pair $(items_T, tids_T)$ is a Galois connection between the posets $(\mathcal{P}(D), \subseteq)$ and $(\mathcal{P}(I), \subseteq)$.*

*Proof.* We need to prove that

1. if $E \subseteq E'$, then $items_T(E') \subseteq items_T(E)$,
2. if $H \subseteq H'$, then $tids_T(H') \subseteq tids_T(H)$,
3. $E \subseteq tids_T(items_T(E))$, and
4. $H \subseteq items_T(tids_T(H))$

for every $E, E' \in \mathcal{P}(D)$ and every $H, H' \in \mathcal{P}(I)$.

The first two properties follow immediately from the definitions of the functions $items_T$ and $tids_T$.

To prove Part (iii), let $k \in E$ be a transaction identifier. Then, the item set $T(e)$ includes $items_T(E)$ by the definition of $items_T(E)$. By Part (ii), $tids_T(T(e)) \subseteq tids_T(items_T(E))$. Since $e \in tids_T(T(e))$, it follows that $e \in tids_T(items_T(E))$, so $E \subseteq tids_T(items_T(E))$.

The argument for Part (iv) is similar. □

The theorem can be obtained directly by noting that $(items_T, tids_T)$ is the polarity determined by the relation

$$\rho = \{(k, i) \in D \times I \mid i \in T(k)\}.$$

**Corollary 7.33.** *Let $T : D \longrightarrow \mathcal{P}(I)$ be a transaction data set and let $\boldsymbol{K}_i : \mathcal{P}(I) \longrightarrow \mathcal{P}(I)$ and $\boldsymbol{K}_d : \mathcal{P}(D) \longrightarrow \mathcal{P}(D)$ be defined by $\boldsymbol{K}_i(H) = items_T(tids_T(H))$ for $H \in \mathcal{P}(I)$ and $\boldsymbol{K}_d(E) = tids_T(items_T(E))$ for $E \in \mathcal{P}(D)$. Then, $\boldsymbol{K}_i$ and $\boldsymbol{K}_d$ are closure operators on $I$ and $D$, respectively.*

*Proof.* The argument was made in Example 5.52. □

**Theorem 7.34.** *Let $T : D \longrightarrow \mathcal{P}(I)$ be a transaction data set. We have*

$$\boldsymbol{K}_i(H_1 \cup H_2) = \boldsymbol{K}_i(H_1) \cap \boldsymbol{K}_i(H_2),$$
$$\boldsymbol{K}_d(E_1 \cup E_2) = \boldsymbol{K}_d(E_1) \cap \boldsymbol{K}_d(E_2),$$

*for $H_1, H_2 \subseteq I$ and $E_1, E_2 \subseteq D$.*

*Proof.* This statement is a direct consequence of the definitions of $\mathbf{K}_i$ and $\mathbf{K}_d$. □

Closed sets of items (that is, sets of items $H$ such that $H = \mathbf{K}_i(H)$) can be characterized as follows.

**Theorem 7.35.** *Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set.*

*A set of items $H$ is closed if and only if, for every set $L \in \mathcal{P}(I)$ such that $H \subset L$, we have $supp_T(L) < supp_T(H)$.*

*Proof.* Suppose that for every superset $L$ of $H$ we have $supp_T(H) > supp_T(L)$ and that $H$ is not a closed set of items. Therefore, the set $\mathbf{K}_i(H) = items_T(tids_T(H))$ is a superset of $H$ and consequently $suppcount_T(H) > suppcount_T(items_T(tids_T(H)))$. Since

$$suppcount_T(items_T(tids_T(H))) = |tids_T(items_T(tids_T(H)))| = |tids_T(H)|,$$

this leads to a contradiction. Thus, $H$ must be closed.

Conversely, suppose that $H$ is a closed set of items,

$$H = \mathbf{K}_i(H) = items_T(tids_T(H)),$$

and let $L$ be a strict superset of $H$. Suppose that $supp_T(L) = supp_T(H)$. This means that $|tids_T(L)| = |tids_T(H)|$.

Since $H = items_T(tids_T(H)) \subset L$, it follows that

$$tids_T(L) \subseteq tids_T(items_T(tids_T(H))) = tids_T(H),$$

which implies the equality $tids_T(L) = tids_T(items_T(tids_T(H)))$ because the sets $tids_T(L)$ and $tids_T(H)$ have the same number of elements. Thus, we have $tids_T(L) = tids_T(H)$. In turn, this yields

$$H = items_T(tids_T(H)) = items_T(tids_T(L)) \supseteq L,$$

which contradicts the initial assumption $H \subset L$. □

The importance of determining the closed item sets is based on the equality $suppcount_T(items_T(tids_T(H))) = |tids_T(items_T(tids_T(H)))| = |tids_T(H)|$. Thus, if we have the support counts of the closed sets, we have the support count of every set of items and the number of closed sets can be much smaller than the total number of item sets. An interesting algorithm focused on closed item sets was developed in [148].

## Exercises and Supplements

1. Let $I = \{a, b, c, d\}$ be a set of items and let $T$ be a transaction data set defined by

$$\begin{aligned} T(1) &= abc, \\ T(2) &= abd, \\ T(3) &= acd, \\ T(4) &= bcd, \\ T(5) &= ab. \end{aligned}$$

   a) Find item sets whose support is at least 0.25.
   b) Find association rules having support at least 0.25 and a confidence at least 0.75.

2. Let $I = i_1i_2i_3i_4i_5$ be a set of items. Find the 0.6-frequent item sets of the transaction data set $T$ on $I$ defined by

$$\begin{array}{ll} T(1) = i_1, & T(6) = i_1i_2i_4, \\ T(2) = i_1i_2, & T(7) = i_1i_2i_5, \\ T(3) = i_1i_2i_3, & T(8) = i_2i_3i_4, \\ T(4) = i_2i_3, & T(9) = i_2i_3i_5, \\ T(5) = i_2i_3i_4, & T(10) = i_3i_4i_5. \end{array}$$

   Also, determine all rules whose confidence is at least 0.75.

3. Let $T$ be a transaction data set $T$ on an item set $I$, $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$. Define the bit sequence of an item set $X$ as sequence $\mathbf{b}^X = (b_1, \ldots, b_n) \in \mathbf{Seq}_n(\{0, 1\})$, where

$$b_i = \begin{cases} 1 & \text{if } X \subseteq T(i), \\ 0 & \text{otherwise,} \end{cases}$$

   for $1 \leq i \leq n$.
   For $\mathbf{b} \in \mathbf{Seq}_n(\{0, 1\})$, the number $\sqrt{|\{i | 1 \leq i \leq n, b_i = 1\}|}$ is denoted by $\| \mathbf{b} \|$. The distance between the sequences $\mathbf{b}$ and $\mathbf{c}$ is defined as $\| \mathbf{b} \oplus \mathbf{c} \|$. Prove that:

a) $\mathbf{b}^{X \cup Y} = \mathbf{b}^X \wedge \mathbf{b}^Y$ for every $X, Y \in \mathcal{P}(I)$.
b) $\mathbf{b}^{K \oplus L} = \mathbf{b}^L \oplus \mathbf{b}^K$, where $K \oplus L$ is the symmetric difference of the item sets $K$ and $L$.
c) $|\sqrt{supp_T(K)} - \sqrt{supp_T(L)}| \leq \frac{d(\mathbf{b}^K, \mathbf{b}^L)}{\sqrt{|T|}}$.

4. For a transaction data set $T$ on an item set $I = \{i_1, \ldots, i_n\}$, $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ and a number $h$, $1 \leq h \leq n$, define the number $\nu_T(h)$ by
$$\nu_T(h) = 2^{n-1} b_n + \cdots + 2b_2 + b_1,$$
where
$$b_k = \begin{cases} 1 & \text{if } i_k \in T(h), \\ 0 & \text{otherwise,} \end{cases}$$
for $1 \leq k \leq n$. Prove that $i_k \in T(h)$ if and only if the result of the integer division $\nu_T(h)/k$ is an odd number.

Suppose that the tabular variables of a database $\mathcal{D}$ are $(T_1, H_1), \ldots, (T_p, H_p)$. An *inclusion dependency* is an expression of the form $T_i[K] \subseteq T_j[L]$, where $K \subseteq H_i$ and $L \subseteq H_j$ for some $i, j$, where $1 \leq i, j \leq p$ are two sets of attributes having the same cardinality. Denote by $\mathrm{ID}_{\mathcal{D}}$ the set of inclusion dependencies of $\mathcal{D}$.

Let $D \in \mathcal{S}_{\mathcal{D}}$ be a state of the database $\mathcal{D}$, $\phi = T_i[K] \subseteq T_j[L]$ be an inclusion dependency, and $\theta_i = (T_i, H_i, \rho_i)$, $\theta_j = (T_j, H_j, \rho_j)$ be the tables that correspond to the tabular variables $(T_i, H_i)$ and $(T_j, H_j)$ in $D$. The inclusion dependency $\phi$ is satisfied in the state $D$ of $\mathcal{D}$ if for every tuple $t \in \rho_i$ there is a tuple $s \in \rho_j$ such that $t[K] = s[L]$.

5. For $\phi = T_i[K] \subseteq T_j[L]$ and $\psi = T_d[K'] \subseteq T_e[L']$, define the relation $\phi \leq \psi$ if $d = i, e = j$, $K \subseteq K'$, and $H \subseteq H'$. Prove that "$\leq$" is a partial order on $\mathrm{ID}_{\mathcal{D}}$.
6. Prove that the triple $(\mathrm{ID}_{\mathcal{D}}, \leq, h)$ is a graded poset, where $h(T_i[K] \subseteq T_j[L]) = |K|$.
7. Prove that the function $q : \mathcal{S}_{\mathcal{D}} \times \mathrm{ID}_{\mathcal{D}} \longrightarrow \{0, 1\}$ defined by
$$q(D, \phi) = \begin{cases} 1 & \text{if } \phi \text{ is satisfied in } D, \\ 0 & \text{otherwise,} \end{cases}$$
is a query (as in Definition 7.25).
8. Specialize the generic levelwise algorithm to an algorithm that retrieves all inclusion dependencies satisfied by a database state.

Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(D)$ be a transaction data set on an item set $D$. The contingency matrix of two item sets $X$ and $Y$ is the $2 \times 2$-matrix:
$$M_{XY} = \begin{pmatrix} m_{11} & m_{10} \\ m_{01} & m_{00} \end{pmatrix},$$

where

$$m_{11} = |\{k|X \subseteq T(k) \text{ and } Y \subseteq T(k)\}|,$$
$$m_{10} = |\{k|X \subseteq T(k) \text{ and } Y \not\subseteq T(k)\}|,$$
$$m_{01} = |\{k|X \not\subseteq T(k) \text{ and } Y \subseteq T(k)\}|,$$
$$m_{00} = |\{k|X \not\subseteq T(k) \text{ and } Y \not\subseteq T(k)\}|.$$

Also, let $m_{1\cdot} = m_{11} + m_{10}$ and $m_{\cdot 1} = m_{11} + m_{01}$.

9. Let $X \Rightarrow Y$ be an association rule. Prove that
$$conf_T(X \Rightarrow Y) = \frac{m_{11}}{m_{11} + m_{10}}.$$
Which significance has the number $m_{10}$ for $X \Rightarrow Y$?
10. Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set on a set of items $I$ and let $\pi$ be a partition of the set $\{1, \ldots, n\}$ of transaction identifiers, $\pi = \{B_1, \ldots, B_p\}$. Let $n_i = |B_i|$ for $1 \leq i \leq p$.
A *partitioning* of $T$ is a sequence $T_1, \ldots, T_p$ of transaction data sets on $I$ such that $T_i : \{1, \ldots, n_i\} \longrightarrow \mathcal{P}(I)$ is defined by $T_i(\ell) = T(k_\ell)$, where $B_i = \{k_1, \ldots, k_{n_i}\}$ for $1 \leq i \leq p$.
Intuitively, this corresponds to splitting horizontally the table of $T$ into $p$ tables that contain $n_1, \ldots, n_p$ consecutive rows, respectively.
Let $K$ be an item set. Prove that if $supp_T(K) \geq \mu$, there exists $j$, $1 \leq j \leq p$, such that $supp_{T_j}(K) \geq \mu$. Give an example to show that the reverse implication does not hold; in other words, give an example of a transaction data set $T$, a partitioning $T_1, \ldots, T_p$ of $T$, and an item set $K$ such that $K$ is $\mu$-frequent in some $T_i$ but not in $T$.
11. Piatetsky-Shapiro formulated in [107] three principles that a rule interestingness measure $R$ should satisfy:
    a) $R(X \Rightarrow Y) = 0$ if $m_{11} = \frac{m_{1\cdot} m_{\cdot 1}}{n}$,
    b) $R(X \rightarrow Y)$ increases with $m_{11}$ when other parameters are fixed, and
    c) $R(X \rightarrow Y)$ decreases with $m_{\cdot 1}$ and with $m_{1\cdot}$ when other parameters are fixed.

    The *lift* of a rule $X \Rightarrow Y$ is the number $lift(X \Rightarrow Y) = \frac{n m_{11}}{m_{1\cdot} m_{\cdot 1}}$. The *PS* measure is $PS(X \rightarrow Y) = m_{11} - \frac{m_{1\cdot} m_{\cdot 1}}{n}$. Do *lift* and *PS* satisfy Piatetsky-Shapiro's principles? Give examples of interestingness measures that satisfy these principles.

## Bibliographical Comments

In addition to general data mining references [132, 130], the reader should consult [1], a monograph dedicated to frequent item sets and association rules. Seminal work in this area, in addition to the original paper [5], has been done by H. Mannila and H. Toivonen [95] and by M. J. Zaki [148]; these references

lead to an interesting and rewarding journey through the data mining literature. An alternative method for detecting frequent item sets based on a very interesting condensed representation of the data set was developed by Jiawei Han et al. [64].

An algorithm that searches the collection of item sets in a depth-first manner with the purpose of discovering maximal frequent item sets was proposed in [2] and [3].

Exercises 5–8 are reformulations of results obtained in [95].

# 8

# Applications to Databases and Data Mining

## 8.1 Introduction

In this chapter, we discuss various applications of partially ordered sets and some of their associated structures (closure operators), the lattice of partitions, monotonicity of functions, etc.

## 8.2 Tables and Indiscernibility Relations

**Definition 8.1.** *Let* $\theta = (T, H, \boldsymbol{r})$ *be a table, where* $\boldsymbol{r} = (t_1, \ldots, t_n)$. *The* indiscernibility relation defined by a set of attributes $X$, $X \subseteq set(H)$ *is the relation* $\epsilon^X \subseteq \{1, \ldots, n\}^2$ *given by*

$$\epsilon^X = \{(p, q) \in \{1, \ldots, n\}^2 \mid t_p[X] = t_q[X]\}.$$

It is easy to verify that $\epsilon^X$ is an equivalence for every set of attributes $X$. The partition of $\{1, \ldots, n\}$ that corresponds to this equivalence will be denoted by $\pi^X$.

*Example 8.2.* Consider again the table introduced in Example 1.153.

OBJECTS

| | shape | length | width | height | color |
|---|---|---|---|---|---|
| 1 | cube | 5 | 5 | 5 | red |
| 2 | sphere | 3 | 3 | 3 | blue |
| 3 | pyramid | 5 | 6 | 4 | blue |
| 4 | cube | 2 | 2 | 2 | red |
| 5 | sphere | 3 | 3 | 3 | blue |

Several partitions defined by sets of attributes of this table are:

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_8, © Springer-Verlag London Limited 2008

$$\begin{aligned}
\pi^{\text{shape}} &= \{\{1,4\},\{2,5\},\{3\}\}\\
\pi^{\text{length}} &= \{\{1,3\},\{2,5\},\{4\}\},\\
\pi^{\text{width}} &= \{\{1\},\{2,5\},\{3\},\{4\}\},\\
\pi^{\text{height}} &= \{\{1\},\{2,5\},\{3\},\{4\}\},\\
\pi^{\text{color}} &= \{\{1,4\},\{2,3,5\}\},\\
\pi^{\text{shape length}} &= \{\{1\},\{4\},\{2,5\},\{3\}\},\\
\pi^{\text{shape color}} &= \{\{1,4\},\{2,5\},\{3\}\}.
\end{aligned}$$

**Theorem 8.3.** *Let $\theta = (T, H, \boldsymbol{r})$ be a table and let $X$ and $Y$ be two sets of attributes, $X, Y \subseteq H$. We have $\epsilon^{XY} = \epsilon^X \cap \epsilon^Y$.*

*Proof.* Let $t_p, t_q \in \mathit{set}(\mathbf{r})$ be two tuples such that $(p,q) \in \epsilon^{XY}$. This means that $t_p[XY] = t_q[XY]$. By the second part of Theorem 1.154, this holds if and only if $t_p[X] = t_p[X]$ and $t_p[Y] = t_q[Y]$; that is, if and only if $(p,q) \in \epsilon^X$ and $(p,q) \in \epsilon^Y$. Thus, $\epsilon^{XY} = \epsilon^X \cap \epsilon^Y$. □

**Corollary 8.4.** *Let $\theta = (T, H, \boldsymbol{r})$ be a table and let $X$ and $Y$ be two sets of attributes, $X, Y \subseteq H$. We have $\pi^{XY} = \pi^X \wedge \pi^Y$.*

*Proof.* This statement follows immediately from Theorems 8.3 and 4.60. □

**Corollary 8.5.** *Let $\theta = (T, H, \boldsymbol{r})$ be a table and let $X$ and $Y$ be two sets of attributes, $X, Y \subseteq H$. If $X \subseteq Y$, we have $\pi^Y \leq \pi^X$.*

*Proof.* Since $X \subseteq Y$, we have $XY = Y$, so $\pi^Y = \pi^X \wedge \pi^Y$, which implies $\pi^Y \leq \pi^X$. □

**Definition 8.6.** *A* reduct *of a table $\theta = (T, H, \boldsymbol{r})$ is a set of attributes $L$ that satisfies the following conditions:*

(i) $\pi^L = \pi^H$, *and*

(ii) *$L$ is a minimal set having the property (i); that is, for every $J \subset L$, we have $\pi^J \geq \pi^H$.*

*The* core *of $\theta$ is the intersection of the reducts of the table.*

*Example 8.7.* Let $\theta = (T, ABCDE, \mathbf{r})$ be the following table:

$T$

| | $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|---|
| 1 | $a_1$ | $b_1$ | $c_1$ | $d_1$ | $e_1$ |
| 2 | $a_2$ | $b_2$ | $c_2$ | $d_2$ | $e_1$ |
| 3 | $a_1$ | $b_2$ | $c_2$ | $d_1$ | $e_2$ |
| 4 | $a_2$ | $b_2$ | $c_1$ | $d_2$ | $e_2$ |
| 5 | $a_1$ | $b_1$ | $c_1$ | $d_1$ | $e_1$ |
| 6 | $a_1$ | $b_1$ | $c_1$ | $d_1$ | $e_1$ |
| 7 | $a_1$ | $b_2$ | $c_2$ | $d_1$ | $e_2$ |

We have $\pi^H = \{\{1,5,6\},\{3,7\},\{2\},\{4\}\}$. Note that we have $\pi^{AC} = \pi^H$ and $\pi^{DE} = \pi^H$. On the other hand, we also have

$$\begin{aligned}\pi^A &= \{\{1,3,5,6,7\},\{2,4\}\},\\ \pi^C &= \{\{1,4,5,6\},\{2,3,7\}\},\\ \pi^D &= \{\{1,3,5,6,7\},\{2,4\}\},\\ \pi^E &= \{\{1,2,5,6\},\{3,4,7\}\},\end{aligned}$$

which shows that both $AC$ and $DE$ are reducts of this table.

Table reducts are minimal sets of attributes that have the same separating power as the entire set of attributes of the table. Example 8.7 shows that a table may possess several reducts.

Note that no two distinct reducts may be comparable as sets because of the minimality condition. Therefore, each maximal chain of sets in the poset $(\mathcal{P}(H), \subseteq)$ that joins $\emptyset$ to $H$ may include at most one reduct. Thus, the largest number of reducts that a table with $n$ attributes may have is $\binom{n}{\lfloor n/2 \rfloor}$.

*Example 8.8.* Let $\theta$ be the table

$T$

| | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| 1 | $a_1$ | $b_1$ | $c_1$ | $d_1$ |
| 2 | $a_1$ | $b_2$ | $c_1$ | $d_2$ |
| 3 | $a_2$ | $b_1$ | $c_1$ | $d_1$ |
| 4 | $a_2$ | $b_2$ | $c_1$ | $d_2$ |

It is easy to see that this table has two reducts, $AB$ and $AD$. Therefore, the core of this table consists of the attribute $A$.

On the other hand, the core of the two-tuple table

$S$

| | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| 1 | $a_1$ | $b_1$ | $c_1$ | $d_1$ |
| 2 | $a_1$ | $b_2$ | $c_1$ | $d_2$ |

is empty because its two reducts, $B$ and $D$, have no attributes in common.

The following theorem gives a characterization of table reducts.

**Theorem 8.9.** *Let $\theta = (T, H, \boldsymbol{r})$ be a table such that $|\boldsymbol{r}| = n$ and let $\delta : \{1,\ldots,n\}^2 \longrightarrow \mathcal{P}(H)$ be the function defined by*

$$\delta(i,j) = \{A \in set(H) \mid t_i[A] \neq t_j[A]\}$$

*for $1 \leq i,j \leq n$. The set of attributes $L$ is a reduct for $\theta$ if and only if $L \cap \delta(i,j) \neq \emptyset$ for every pair $(i,j) \in \{1,\ldots,n\}^2$ such that $\delta(i,j) \neq \emptyset$ and $L$ is minimal with this property.*

*Proof.* Suppose that $L$ is a reduct for $\theta$ and that $(i, j)$ is a pair such that $\delta(i, j) \neq \emptyset$. The equality $\delta(i, j) \neq \emptyset$ implies that $(i, j) \notin \epsilon^H = \epsilon^L$, so $t_i[L] \neq t_j[L]$. Therefore, $L \cap \delta(i, j) \neq \emptyset$.

Suppose that there is a strict subset $G$ of $L$ such that $G \cap \delta(i, j) \neq \emptyset$ for every pair $(i, j) \in \{1, \ldots, n\}^2$ such that $\delta(i, j) \neq \emptyset$. This implies $\epsilon^G = \epsilon^H$, which contradicts the minimality of the reduct $L$.

Conversely, suppose that $L \cap \delta(i, j) \neq \emptyset$ for every pair $(i, j) \in \{1, \ldots, n\}^2$ such that $\delta(i, j) \neq \emptyset$ and $L$ is minimal with this property. Since $L \subseteq H$, we have $\epsilon^H \subseteq \epsilon^L$.

Now let $(h, k)$ be a pair in $\epsilon^L$. Since $t_h$ coincides with $t_k$ on every attribute of $L$, it follows that we must have $\delta(h, k) = \emptyset$, which implies $(h, k) \in \epsilon^H$. Thus, $\epsilon^H = \epsilon^L$. If $L$ is a minimal set satisfying the condition of the theorem, it follows immediately that $L$ is minimal in the collection of sets of attributes that differentiate the tuples of $\theta$, so $L$ is a reduct. □

The notion of a key that is frequently used in databases is related to the notion of a reduct.

**Definition 8.10.** *A* key *of a table $\theta = (T, H, \boldsymbol{r})$ and $\boldsymbol{r} = (t_1, \ldots, t_n)$ is a set of attributes $L$ that satisfies the following conditions:*
(i) $\pi^L = \alpha_{\{1,\ldots,n\}}$, *and*
(ii) *$L$ is a minimal set having the property (i); that is, for every $J \subset L$, we have $\pi^J \geq \alpha_{\{1,\ldots,n\}}$.*

A table $\theta = (T, H, \mathbf{r})$ has a key if and only if the sequence $\mathbf{r}$ does not contain duplicate tuples.

## 8.3 Partitions and Functional Dependencies

If we examine the table defined in Example 8.2, we observe that if two objects have the same shape, then they have the same color. We also note that the reverse implication is not true because two objects may have the same color without having the same shape. This observation suggests the introduction of a type of constraint that applies to the table contents for every table that is a value of a tabular variable.

**Definition 8.11.** *Let $H$ be a set of attributes. A* functional dependency *is an ordered pair $(X, Y)$ of subsets of $H$.*

The set of all functional dependencies on a set of attributes $H$ is denoted by $\mathsf{FD}(H)$. If $(X, Y) \in \mathsf{FD}(H)$ we shall write this pair as $X \to Y$ using a well-established convention in database theory.

**Definition 8.12.** *Let $\theta = (T, H, \boldsymbol{r})$ be a table and let $X$ and $Y$ be two subsets of $H$. The table $\theta$* satisfies the functional dependency $X \longrightarrow Y$ *if $u[X] = v[X]$ implies $u[Y] = v[Y]$ for every two tuples $u, v \in set(\boldsymbol{r})$.*

In other words, a table $\theta$ satisfies the functional dependency $X \to Y$ if and only if $\epsilon^X \subseteq \epsilon^Y$ or, equivalently, $\pi^X \leq \pi^Y$.

*Example 8.13.* Let us consider a tabular variable whose values are intended to store the data reflecting the instructors, students, and musical instruments studied by the students of a community music school. Lessons are scheduled once a week, and each instructor is teaching one instrument:

$$\tau = (\text{SCHEDULE}, \text{student instructor instrument day time room}).$$

Any table $\theta$ that is a value of this tabular variable must satisfy functional dependencies that reflect these "business rules" as well as other semantic restrictions:

student instrument → instructor,
instructor → instrument,
student instrument → day time,
room day time → student instructor,
student day time → room,
instructor day time → room.

For example, a possible value of this tabular variable is the table:

SCHEDULE

| | student | instructor | instrument | day | time | room |
|---|---|---|---|---|---|---|
| 1 | Margo | Donna | piano | Mon | 4 | A |
| 2 | Danielle | Igor | violin | Mon | 4 | B |
| 3 | Joshua | Donna | piano | Mon | 5 | A |
| 4 | Ondine | Donna | piano | Tue | 3 | A |
| 5 | Michael | Donna | piano | Tue | 4 | A |
| 6 | Linda | Mary | flute | Tue | 4 | B |
| 7 | Todor | Mary | flute | Tue | 5 | A |
| 8 | Sarah | Emma | piano | Tue | 6 | A |
| 9 | Samuel | Donna | piano | Tue | 6 | B |
| 10 | Alex | David | guitar | Tue | 6 | C |
| 11 | Dan | Emma | piano | Wed | 3 | A |
| 12 | William | Mary | flute | Wed | 4 | A |
| 13 | Nora | David | guitar | Wed | 4 | B |
| 14 | Amy | Donna | piano | Wed | 5 | A |
| 15 | Peter | Igor | violin | Thr | 4 | A |
| 16 | Kenneth | David | guitar | Thr | 4 | B |
| 17 | Patrick | Donna | piano | Thr | 5 | A |
| 18 | Elizabeth | Emma | piano | Thr | 5 | B |
| 19 | Helen | Mary | flute | Thr | 5 | C |
| 20 | Cris | Mary | flute | Fri | 4 | B |
| 21 | Richard | Igor | violin | Fri | 4 | C |
| 22 | Yves | Donna | piano | Fri | 5 | A |
| 23 | Paul | Emma | piano | Fri | 5 | B |
| 24 | Colin | Igor | violin | Fri | 6 | C |

The reader can easily check that this table satisfies all functional dependencies identified after the definition of the tabular variable.

It is clear that if $X, Y \subseteq H$ and $Y \subseteq X$, any table $\theta = (T, H, \mathbf{r})$ satisfies the functional dependency $X \to Y$.

**Definition 8.14.** *A functional dependency $X \to Y$ is* trivial *if it is satisfied by every table $\theta = (T, H, \mathbf{r})$ such that $X, Y \in \mathcal{P}(H)$.*

**Theorem 8.15.** *Let $H$ be a finite set of attributes. A functional dependency $X \to Y \in \mathsf{FD}(H)$ is trivial if and only if $Y \subseteq X$.*

*Proof.* For each attribute $A \in H$, let $u_A$ and $v_A$ be two distinct values in $\mathrm{Dom}(A)$. Suppose that $X \to Y$ is a trivial functional dependency and that $Y$ is not included in $X$. This means that there exists an attribute $B \in Y - X$. Consider the table $\theta = (T, XY, \mathbf{r})$, where $\mathbf{r} = (t_1, t_2)$, where $t_1[A] = u_A$ for every $A \in XY$, and

$$t_2[A] = \begin{cases} u_A & \text{if } A \neq B, \\ v_B & \text{if } A = B. \end{cases}$$

Since $t_1[X] = t_2[X]$ and $t_1[Y] \neq t_2[Y]$, it follows that $\theta$ violates the functional dependency $X \to Y$, which contradicts the fact that $X \to Y$ is trivial.

The sufficiency of the condition is immediate. □

Suppose now that $\theta = (T, H, \mathbf{r})$ satisfies the functional dependencies $X \to Y$ and $Y \to Z$, where $X, Y, Z$ are subsets of $H$. This means that $\pi^X \leq \pi^Y$ and $\pi^Y \leq \pi^Z$, which implies $\pi^X \leq \pi^Z$. Therefore, $\theta$ satisfies the functional dependency $X \to Z$.

If $\theta = (T, H, \mathbf{r})$ satisfies the functional dependency $X \to Y$ and $W$ is a subset of $H$, then we have $\pi^X \leq \pi^Y$. Therefore, we have

$$\pi^{XW} = \pi^X \wedge \pi^W \leq \pi^Y \wedge \pi^W = \pi^{YW},$$

which means that $\theta$ satisfies the functional dependency $XW \to YW$.

In the database design process, it is necessary to identify functional dependencies satisfied by tables that are values of tabular variables. Thus, for a tabular variable $\tau = (T, H)$, the design of the database entails the construction of the *functional dependency schema* defined as a pair $S = (H, F)$, where $F \subseteq \mathsf{FD}(H)$. Tables that are values of $\tau$ are also said to *satisfy the schema $S$*. The identification of these functional dependencies is based on the meaning of the attributes involved. For example, in a table schema that contains the attributes *ssn* (standing for social security number) and *name*, it is natural to impose the functional dependency *ssn* → *name*. Every table that satisfies this schema will satisfy this functional dependency.

Suppose that a table satisfies the functional dependencies $A \to B$ and $B \to C$. Then, by a previous observation, the table will also satisfy $A \to C$. Thus, it is not necessary to explicitly stipulate that the table will satisfy $A \to C$. This functional dependency is obtained by applying the rule

$$\frac{A \to B, B \to C}{A \to C},$$

which is an instance of the *transitivity rule*

$$\frac{X \to Y, Y \to Z}{X \to Z}\, R_{trans},$$

for every $X, Y, Z \in \mathcal{P}(H)$. Here $H$ is the set of attributes of a table. Our previous argument shows that this rule is sound; in other words, if a table satisfies $X \to Y$ and $Y \to Z$, then the table satisfies $X \to Z$.

The previous arguments allow us to identify two more sound rules, the *inclusion rule*

$$\frac{X \subseteq Y}{Y \to X} R_{inc},$$

and the *augmentation rule*

$$\frac{X \to Y}{XW \to YW} R_{aug},$$

for every $X, Y, W \in \mathcal{P}(H)$.

As we saw above, rules are denoted as fractions; the objects that appear in the numerator are known as the *premises* of the rule; the object that appears in the denominator is the *conclusion* of the rule.

The three rules introduced so far (transitivity, augmentation, and inclusion) are known as *Armstrong's rules*.

The previous discussion establishes the *soundness* of the rules $R_{inc}$, $R_{aug}$, and $R_{tran}$. This means that a table that satisfies a set $F$ of functional dependencies will satisfy any functional dependency obtained from $F$ through applications of these rules.

In a certain sense that will be made clear in what follows, these are all the rules we need in order to reason about functional dependencies.

Rules are used to generate in a syntactic manner new functional dependencies starting from existing sets of such dependencies. The process of producing such new functional dependencies is known as a *proof*. This notion is formalized in the next definition.

**Definition 8.16.** *Let $S = (H, F)$ be a functional dependencies schema. A non-null sequence of functional dependencies:*

$$U_1 \to V_1, \ldots, U_n \to V_n$$

*is an $F$-*proof of length *$n$ if one of the following conditions is satisfied for every $i$, $1 \leq i \leq n$:*

(i) *$U_i \to V_i$ is one of the functional dependencies of $F$, or*

(ii) *$U_i \to V_i$ is obtained from $0, 1$, or $2$ predecessors in the sequence by applying one of Armstrong's rules.*

*The last dependency in the sequence $U_n \to V_n$ is the* target *of the $F$-proof.*

*Example 8.17.* Suppose that $S = (H, F)$ is a functional dependency schema, where $H = ABCDE$ and $F$ is the set of functional dependencies

$$F = \{A \to C, AB \to D, CD \to E\}.$$

We claim that the sequence

$$A \to C, AB \to BC, AB \to ABC, AB \to D,$$
$$ABC \to CD, AB \to CD, CD \to E, AB \to E$$

is an $F$-proof of $AB \to E$ for the following reasons:
(i) $A \to C$ belongs to $F$.
(ii) $AB \to ABC$ is obtained from (i) by applying $R_{aug}$ with $W = AB$.
(iii) $AB \to D$ belongs to $F$.
(iv) $ABC \to CD$ is obtained from (iii) by applying $R_{aug}$ with $W = C$.
(v) $AB \to CD$ is obtained from (ii) and (iv) by applying $R_{tran}$.
(vi) $CD \to E$ belongs to $F$.
(vii) $AB \to E$ is obtained from (v) and (vi) by applying $R_{tran}$.

The existence of an $F$-proof that has a functional dependency $U \to V$ as a target is denoted as $F \underset{ARM}{\vdash} U \to V$.

Finding an $F$-proof for a functional dependency can be a daunting task if the number of attributes and functional dependencies is large. Fortunately, there are ways of simplifying this process.

**Theorem 8.18.** *Let $S = (H, F)$ be a functional dependency schema. If $F \underset{ARM}{\vdash} X \to Y$ and $F \underset{ARM}{\vdash} X \to Z$, then $F \underset{ARM}{\vdash} X \to YZ$ for every $X, Y, Z \subseteq H$.*

*Proof.* Let $U_1 \to V_1, \ldots, U_n \to V_n$ and $U'_1 \to V'_1, \ldots, U'_n \to V'_m$ be two $F$-proofs that have $X \to Y$ and $X \to Z$ as targets. Using the augmentation by $X$, the first proof generates the $F$-proof

$$U_1 \to V_1, \ldots, U_n \to V_n = X \to Y, X \to XY.$$

On the other hand, starting from the second proof

$$U'_1 \to V'_1, \ldots, U'_n \to V'_m = X \to Z,$$

by augmenting the last functional dependency by $Y$, we have the $F$-proof

$$U'_1 \to V'_1, \ldots, U'_n \to V'_m = X \to Z, XY \to YZ$$

By concatenating the two newly obtained proofs and applying the transitivity property, we have the $F$-proof

$$U_1 \to V_1, \ldots, U_n \to V_n, X \to XY,$$
$$U'_1 \to V'_1, \ldots, U'_n \to V'_m, XY \to YZ, X \to YZ,$$

which has the desired functional dependency $X \to YZ$ as its target. □

The last theorem shows that we can derive the functional dependency $X \to YZ$ from the functional dependencies $X \to Y$ and $X \to Z$. This fact is interpreted as a "derived rule" known as the *additivity rule* and is denoted by

$$\frac{X \to Y, X \to Z}{X \to YZ}\ R_{add}.$$

Another derived rule is introduced in the next theorem.

**Theorem 8.19.** *If* $F \underset{ARM}{\vdash} X \to YZ$, *then* $F \underset{ARM}{\vdash} X \to Y$ *for every* $X, Y, Z \subseteq H$.

*Proof.* Let $U_1 \to V_1, \ldots, U_n \to V_n$ be an $F$-proof that has $X \to YZ$ as its target. We can add to this proof the functional dependency $YZ \to Y$ obtained by applying $R_{inc}$. This yields the needed $F$-proof

$$U_1 \to V_1, \ldots, U_n \to V_n = X \to YZ, YZ \to Y, X \to Y,$$

where the last step was obtained by applying the transitivity rule to the previous two steps. □

Thus, from $X \to YZ$ we can derive the functional dependency $X \to Y$. This derived rule is known as the *projectivity rule* and is denoted by

$$\frac{X \to YZ}{X \to Y}\ R_{proj}.$$

Note that if $F$ is a set of functional dependencies, $F \subseteq \mathsf{FD}(H)$ and $X \subseteq H$, then it is always possible to find $Y$ such that $F \underset{ARM}{\vdash} X \to Y$. Indeed, it suffices to take $Y = X$ and we can always prove $X \to X$ starting from $F$ because the functional dependency $X \to X$ can be generated by applying $R_{inc}$.

Let $X \to Y_1, \ldots, X \to Y_p$ be the set of all functional dependencies such that $F \underset{ARM}{\vdash} X \longrightarrow Y$, where $X, Y \subseteq H$. By repeatedly applying the additivity rule, we have $F \underset{ARM}{\vdash} X \longrightarrow Y_1 \cdots Y_p$. The set $Y_1 \cdots Y_p$ is the largest set $Y$ such that $F \underset{ARM}{\vdash} X \longrightarrow Y$. Further, we have $F \underset{ARM}{\vdash} X \longrightarrow V$ if and only if $V \subseteq Y_1 \cdots Y_p$. Indeed, it is clear that if $F \underset{ARM}{\vdash} X \longrightarrow V$, then $V \subseteq Y_1 \cdots Y_p$. Conversely, if $V \subseteq Y_1 \cdots Y_p$, then $Y_1 \cdots Y_p \to V$ (by $R_{inc}$), so $F \underset{ARM}{\vdash} X \to V$ by $R_{tran}$. Thus, the set $Y_1 \cdots Y_p$ plays a special role; we will refer to it as the *closure of* $X$ *under* $F$ and will denote it by $cl_F(X)$.

**Theorem 8.20.** *Let* $S = (H, F)$ *be a functional dependency schema. The mapping* $\mathsf{cl}_F : \mathcal{P}(H) \longrightarrow \mathcal{P}(H)$ *is a closure operator on* $H$.

*Proof.* We need to show that $\mathsf{cl}_F$ satisfies the conditions of Definition 4.37.

Since we have $F \underset{ARM}{\vdash} X \to X$, it is clear that $X \subseteq cl_F(X)$.

Suppose now that $X, X' \in \mathcal{P}(H)$ and $X' \subseteq X$. Since $F \underset{ARM}{\vdash} X \to X'$ by $R_{inc}$ and $F \underset{ARM}{\vdash} X' \to cl_F(X')$, it follows that $F \underset{ARM}{\vdash} X \to cl_F(X')$. This implies $cl_F(X') \subseteq cl_F(X)$, so $\mathsf{cl}_F$ is monotonic.

Finally, note that we have both $F \underset{ARM}{\vdash} X \to cl_F(X)$ and $F \underset{ARM}{\vdash} cl_F(X) \to cl_F(cl_F(X))$, which yields $F \underset{ARM}{\vdash} X \to cl_F(cl_F(X))$ by $R_{tran}$. This implies $cl_F(cl_F(X)) \subseteq cl_F(X)$. The converse inclusion follows from the fact that $X \subseteq cl_F(X)$ and the monotonicity of $\mathsf{cl}_F$. Thus, $cl_F(cl_F(X)) = cl_F(X)$ for every $X \in \mathcal{P}(H)$. □

The statement that $F \underset{ARM}{\vdash} X \to Y$ has a syntactic character; it can be shown by constructing an $F$-proof that has $X \to Y$ as its target. Actually, a computation of $cl_F(X)$ allows us to decide whether $F \underset{ARM}{\vdash} X \to Y$ without constructing the $F$-proof, as shown in the next theorem.

**Theorem 8.21.** *Let $S = (H, F)$ be a functional dependency schema. We have $F \underset{ARM}{\vdash} X \to Y$ if and only if $Y \subseteq cl_F(X)$.*

*Proof.* If $Y \subseteq cl_F(X)$, then we have $F \underset{ARM}{\vdash} cl_F(X) \to Y$ by a single application of $R_{inc}$. Then, since $F \underset{ARM}{\vdash} X \to cl_F(X)$, (by the definition of $cl_F(X)$), another application of $R_{tran}$ yields $F \underset{ARM}{\vdash} X \to Y$.

Conversely, if $F \underset{ARM}{\vdash} X \to Y$, then $Y \subseteq cl_F(X)$ by the definition of $cl_F(X)$. □

Now we introduce a semantic counterpart of relation $\underset{ARM}{\vdash}$.

**Definition 8.22.** *Let $S = (H, F)$ be a functional dependency schema. The set $F$* logically implies *the functional dependency $X \to Y$ if every table that satisfies all functional dependencies of $F$ also satisfies $X \to Y$. This is denoted by $F \models X \to Y$.*

The soundness of Armstrong's rules means that if $F \underset{ARM}{\vdash} X \to Y$, then $F \models X \to Y$. It is interesting that the reverse implication also holds. This fact is known as the *completeness* of Armstrong's axioms and will be established next. To this end, we introduce the notion of an *Armstrong table*.

**Definition 8.23.** *Let $S = (H, F)$ be a functional dependency schema, where $H = A_1 \cdots A_n$, and let $X \in \mathcal{P}(H)$. For each attribute $A \in H$, let $u_A$ and $v_A$ be two distinct values from* Dom$(A)$. *The* Armstrong table *$\theta_{F,X} = (T_{F,X}, H, \boldsymbol{r}_{F,X})$ contains a two-row sequence $\boldsymbol{r}_{F,X}) = (t_0, t_1)$, where $t_0(A) = u_A$ for $A \in H$ and*

$$t_1(A) = \begin{cases} u_A & \text{if } A \in cl_F(X), \\ v_A & \text{if } A \in H - cl_F(X). \end{cases}$$

Note that the existence of Armstrong relations is assured by our assumption that the domain of every attribute contains at least two values.

**Lemma 8.24.** *Let $F$ be a set of functional dependencies $F \subseteq \mathsf{FD}(H)$, $H = A_1 \cdots A_n$, and let $X \in \mathcal{P}(H)$. The* Armstrong table $\theta_{F,X} = (T_{F,X}, H, \boldsymbol{r}_{F,X})$ *satisfies all dependencies that can be proven from $F$.*

*Proof.* Suppose that $U \to V$ is a functional dependency that can be proven from $F$ (which means that $F \underset{ARM}{\vdash} U \to V$) and that this dependency is violated by $\theta_{F,X}$. Since $\theta_{F,X}$ contains two tuples, this is possible only if these tuples have the same projection on $U$ but distinct projections on $V$. The definition of $\theta_{F,X}$ allows this only if $U \subseteq cl_F(X)$ and $V \not\subseteq cl_F(X)$. By the definition of $cl_F(X)$ this is possible only if $F \underset{ARM}{\vdash} X \to U$ and $F \underset{ARM}{\nvdash} X \to V$. This leads to a contradiction because $F \underset{ARM}{\vdash} X \to U$ and $F \underset{ARM}{\vdash} U \to V$ imply $F \underset{ARM}{\vdash} X \to V$ (by $R_{tran}$). □

**Theorem 8.25 (Completeness of Armstrong's Rules).** *Let $F$ be a set of functional dependencies, $F \subseteq \mathsf{FD}(H)$, $H = A_1 \cdots A_n$, and let $X, Y \in \mathcal{P}(H)$. If $F \models X \to Y$, then $F \underset{ARM}{\vdash} X \to Y$.*

*Proof.* Suppose that $F \models X \to Y$ but $F \underset{ARM}{\nvdash} X \to Y$, which means that $Y \not\subseteq cl_F(X)$. The Armstrong table $\theta_{F,X} = (T_{F,X}, H, \mathbf{r}_{F,X})$ satisfies $X \to Y$ because it satisfies all functional dependencies of $F$. Since $X \subseteq cl_F(X)$, this implies $Y \subseteq cl_F(X)$, which yields a contradiction. □

**Corollary 8.26.** *Let $S = (H, F)$ be a functional dependency schema and let $X \to Y$ be a functional dependency in $\mathsf{FD}(F)$. The following three statements are equivalent:*
(i) $Y \subseteq cl_F(X)$.
(ii) $F \underset{ARM}{\vdash} X \to Y$.
(iii) $F \models X \to Y$.

*Proof.* (i) is equivalent to (ii) by Theorem 8.21. We have (ii) implies (iii) by the soundness of Armstrong's rules and (iii) implies (ii) by the completeness of these rules. □

## 8.4 Partition Entropy

The notion of entropy is as a probabilistic concept that lies at the foundation of information theory. Our goal is to define entropy in an algebraic setting by introducing the notion of entropy of a partition taking advantage of the partial order that is naturally defined on the set of partitions of a set. Actually,

we will introduce a generalization of the notion of entropy that has the Gini index and Shannon entropy as special cases.

Let $S$ be a finite set and let $\pi = \{B_1, \ldots, B_m\}$ be a partition of $S$. The *Shannon entropy of* $\pi$ is the number

$$\mathcal{H}(\pi) = -\sum_{i=1}^{m} \frac{|B_i|}{|S|} \log_2 \frac{|B_i|}{|S|}.$$

The *Gini index of* $\pi$ is the number

$$gini(\pi) = 1 - \sum_{i=1}^{m} \left( \frac{|B_i|}{|S|} \right)^2.$$

Both numbers can be used to evaluate the uniformity of the distribution of the elements of $S$ in the blocks of $\pi$ because both values increase with the uniformity of the distribution of the elements of $S$.

*Example 8.27.* Let $S$ be a set containing ten elements and let $\pi_1, \pi_2, \pi_3, \pi_4$ be the four partitions shown in Figure 8.1.

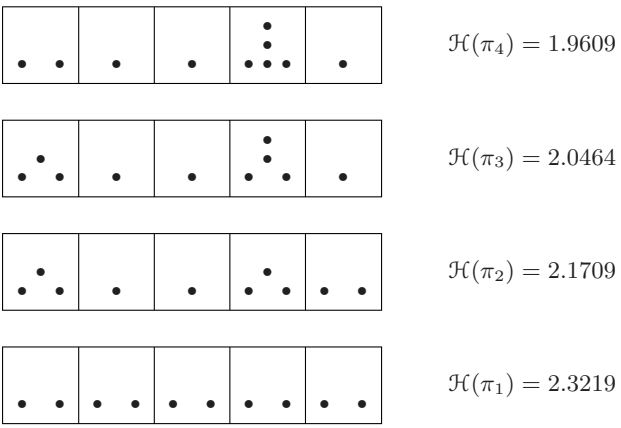



**Fig. 8.1.** Entropy increasing with partition uniformity.

The partition $\pi_1$, which is the most uniform (each block containing two elements), has the largest entropy. At the other end of the range, partition $\pi_4$ has a strong concentration of elements in its fourth block and the lowest entropy. Similar results involving the Gini index are shown in Figure 8.2.

If $S$ and $T$ are two disjoint and nonempty sets, $\pi \in PART(S)$ and $\sigma \in PART(T)$, where $\pi = \{B_1, \ldots, B_m\}$, $\sigma = \{C_1, \ldots, C_n\}$, then the partition $\pi + \sigma$ is the partition of $S \cup T$ given by

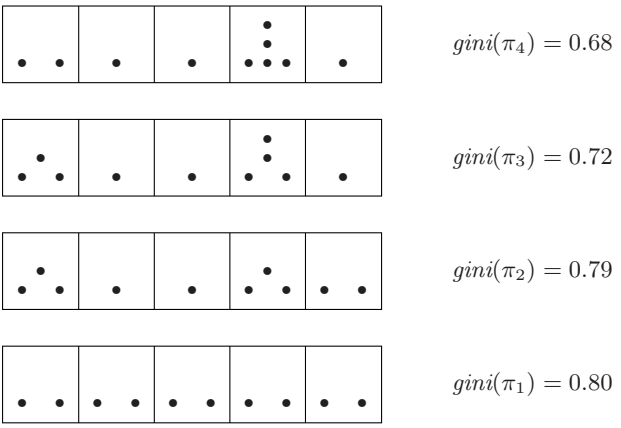



**Fig. 8.2.** Gini index increasing with partition uniformity.

$$\pi + \sigma = \{B_1, \ldots, B_m, C_1, \ldots, C_n\}.$$

Whenever the "+" operation is defined, then it is easily seen to be associative. In other words, if $S, T, U$ are pairwise disjoint and nonempty sets and $\pi \in PART(S)$, $\sigma \in PART(T)$, $\tau \in PART(U)$, then $\pi + (\sigma + \tau) = (\pi + \sigma) + \tau$. Observe that if $S$ and $T$ are disjoint, then $\alpha_S + \alpha_T = \alpha_{S \cup T}$. Also, $\omega_S + \omega_T$ is the partition $\{S, T\}$ of the set $S \cup T$.

If $\pi = \{B_1, \ldots, B_m\}$, $\sigma = \{C_1, \ldots, C_n\}$ are partitions of two arbitrary sets $S, T$, then we denote the partition $\{B_i \times C_j \mid 1 \leq i \leq m, 1 \leq j \leq n\}$ of $S \times T$ by $\pi \times \sigma$. Note that $\alpha_S \times \alpha_T = \alpha_{S \times T}$ and $\omega_S \times \omega_T = \omega_{S \times T}$.

We introduce below a system of four axioms.

**Definition 8.28.** *Let $\beta \in \mathbb{R}$, $\beta \geq 1$, and $\Phi : \mathbb{R}^2_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ be a continuous function such that $\Phi(x, y) = \Phi(y, x)$, and $\Phi(x, 0) = x$ for $x, y \in \mathbb{R}_{\geq 0}$.*

*A $(\Phi, \beta)$-system of axioms for a partition entropy $\mathcal{H}_\beta : PART(S) \longrightarrow \mathbb{R}_{\geq 0}$ consists of the following axioms:*

**(P1)** *If $\pi, \pi' \in PART(S)$ are such that $\pi \leq \pi'$, then $\mathcal{H}_\beta(\pi') \leq \mathcal{H}_\beta(\pi)$.*

**(P2)** *If $S$ and $T$ are two finite sets such that $|S| \leq |T|$, then $\mathcal{H}_\beta(\alpha_S) \leq \mathcal{H}_\beta(\alpha_T)$.*

**(P3)** *For all disjoint sets $S$ and $T$ and partitions $\pi \in PART(S)$ and $\sigma \in PART(T)$ we have*

$$\mathcal{H}_\beta(\pi + \sigma) = \left(\frac{|S|}{|S| + |T|}\right)^\beta \mathcal{H}_\beta(\pi) + \left(\frac{|T|}{|S| + |T|}\right)^\beta \mathcal{H}_\beta(\sigma) + \mathcal{H}_\beta(\{S, T\}).$$

**(P4)** *We have*

$$\mathcal{H}_\beta(\pi \times \sigma) = \Phi(\mathcal{H}_\beta(\pi), \mathcal{H}_\beta(\sigma))$$

*for* $\pi \in PART(S)$ *and* $\sigma \in PART(T)$.

Observe that we postulate that $\mathcal{H}_\beta(\pi) \geq 0$ for any partition $\pi$ since the range of every function $\mathcal{H}_\beta$ is $\mathbb{R}_{\geq 0}$.

**Lemma 8.29.** *For every* $(\Phi, \beta)$*-entropy* $\mathcal{H}_\beta$ *and set* $S$*, we have* $\mathcal{H}_\beta(\omega_S) = 0$.

*Proof.* Let $S$ and $T$ be two disjoint sets that have the same cardinality, $|S| = |T|$. Since $\omega_S + \omega_T$ is the partition $\{S, T\}$ of the set $S \cup T$, by Axiom (P3) we have

$$\mathcal{H}_\beta(\omega_S + \omega_T) = \left(\frac{1}{2}\right)^\beta (\mathcal{H}_\beta(\omega_S) + \mathcal{H}_\beta(\omega_T)) + \mathcal{H}_\beta(\{S, T\}),$$

which implies $\mathcal{H}_\beta(\omega_S) + \mathcal{H}_\beta(\omega_T) = 0$. Since $\mathcal{H}_\beta(\omega_S) \geq 0$ and $\mathcal{H}_\beta(\omega_T)) \geq 0$, it follows that $\mathcal{H}_\beta(\omega_S) = \mathcal{H}_\beta(\omega_T) = 0$. ∎

**Lemma 8.30.** *Let* $S$ *and* $T$ *be two disjoint sets and let* $\pi, \pi' \in PART(S \cup T)$ *be defined by* $\pi = \sigma + \alpha_T$ *and* $\pi' = \sigma + \omega_T$*, where* $\sigma \in PART(S)$*. Then,*

$$\mathcal{H}_\beta(\pi) = \mathcal{H}_\beta(\pi') + \left(\frac{|T|}{|S| + |T|}\right)^\beta \mathcal{H}_\beta(\alpha_T).$$

*Proof.* By Axiom (P3), we can write

$$\begin{aligned}\mathcal{H}_\beta(\pi) = &\left(\frac{|S|}{|S| + |T|}\right)^\beta \mathcal{H}_\beta(\sigma) \\ &+ \left(\frac{|T|}{|S| + |T|}\right)^\beta \mathcal{H}_\beta(\alpha_T) + \mathcal{H}_\beta(\{S, T\})\end{aligned}$$

and

$$\begin{aligned}\mathcal{H}_\beta(\pi') = &\left(\frac{|S|}{|S| + |T|}\right)^\beta \mathcal{H}_\beta(\sigma) \\ &+ \left(\frac{|T|}{|S| + |T|}\right)^\beta \mathcal{H}_\beta(\omega_T) + \mathcal{H}_\beta(\{S, T\}) \\ = &\left(\frac{|S|}{|S| + |T|}\right)^\beta \mathcal{H}_\beta(\sigma) + \mathcal{H}_\beta(\{S, T\}) \\ &\text{(by Lemma 8.29).}\end{aligned}$$

The equalities above immediately imply the equality of the lemma. ∎

**Theorem 8.31.** *For every* $(\Phi, \beta)$*-entropy and partition* $\pi = \{B_1, \ldots, B_m\} \in PART(S)$*, we have*

$$\mathcal{H}_\beta(\pi) = \mathcal{H}_\beta(\alpha_S) - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\alpha_{B_i}).$$

*Proof.* Starting from the partition $\pi$, consider the following sequence of partitions in $PART(S)$:

$$\begin{aligned}
\pi_0 &= \omega_{B_1} + \omega_{B_2} + \omega_{B_3} + \cdots + \omega_{B_m} \\
\pi_1 &= \alpha_{B_1} + \omega_{B_2} + \omega_{B_3} + \cdots + \omega_{B_m} \\
\pi_2 &= \alpha_{B_1} + \alpha_{B_2} + \omega_{B_3} + \cdots + \omega_{B_m} \\
&\vdots \\
\pi_n &= \alpha_{B_1} + \alpha_{B_2} + \alpha_{B_3} + \cdots + \alpha_{B_m}.
\end{aligned}$$

Let $\sigma_j = \alpha_{B_1} + \cdots + \alpha_{B_j} + \omega_{B_{i+2}} + \cdots + \omega_{B_m}$. Then, $\pi_i = \sigma_i + \omega_{B_{i+1}}$ and $\pi_{i+1} = \sigma_i + \alpha_{B_{i+1}}$; therefore, by Lemma 8.30, we have

$$\mathcal{H}_\beta(\pi_{i+1}) = \mathcal{H}_\beta(\pi_i) + \left(\frac{|B_{i+1}|}{|S|}\right)^\beta \mathcal{H}_\beta(\alpha_{B_{i+1}})$$

for $0 \leq i \leq m-1$.

A repeated application of this equality yields

$$\mathcal{H}_\beta(\pi_m) = \mathcal{H}_\beta(\pi_0) + \sum_{i=0}^{m-1} \left(\frac{|B_{i+1}|}{|S|}\right)^\beta \mathcal{H}_\beta(\alpha_{B_{i+1}}).$$

Observe that $\pi_0 = \pi$ and $\pi_m = \alpha_S$. Consequently,

$$\mathcal{H}_\beta(\pi) = \mathcal{H}_\beta(\alpha_S) - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\alpha_{B_i}).$$

∎

Note that if $S$ and $T$ are two sets such that $|S| = |T| > 0$, then, by Axiom (P2), we have $\mathcal{H}_\beta(\alpha_S) = \mathcal{H}_\beta(\alpha_T)$. Therefore, the value of $\mathcal{H}_\beta(\alpha_S)$ depends only on the cardinality of $S$, and there exists a function $\mu : \mathbb{N}_1 \longrightarrow \mathbb{R}_{\geq 0}$ such that $\mathcal{H}_\beta(\alpha_S) = \mu(|S|)$ for every nonempty set $S$. Axiom (P2) also implies that $\mu$ is an increasing function. We will refer to $\mu$ as the *kernel* of the $(\Phi, \beta)$-system of axioms.

**Corollary 8.32.** *Let $\mathcal{H}_\beta$ be a $(\Phi, \beta)$-entropy. For the kernel $\mu$ defined in accordance with Axiom (P2) and every partition $\pi = \{B_1, \ldots, B_m\} \in PART(S)$, we have*

$$\mathcal{H}_\beta(\pi) = \mu(|S|) - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta \mu(|B_i|). \tag{8.1}$$

*Proof.* The statement is an immediate consequence of Theorem 8.31. ∎

**Theorem 8.33.** *Let $\pi = \{B_1, \ldots, B_m\}$ be a partition of the set $S$. Define the partition $\pi'$ obtained by fusing the blocks $B_1$ and $B_2$ of $\pi$ as $\pi' = \{B_1 \cup B_2, B_3, \ldots, B_m\}$ of the same set. Then*

$$\mathcal{H}_\beta(\pi) = \mathcal{H}_\beta(\pi') + \left(\frac{|B_1 \cup B_2|}{|S|}\right)^\beta \mathcal{H}_\beta(\{B_1, B_2\}).$$

*Proof.* A double application of Corollary 8.32 yields

$$\mathcal{H}_\beta(\pi') = \mu(|S|) - \left(\frac{|B_1 \cup B_2|}{|S|}\right)^\beta \mu(|B_1 \cup B_2|) - \sum_{i>2}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta \mu(|B_i|)$$

and

$$\mathcal{H}_\beta(\{B_1, B_2\}) = \mu(|B_1 \cup B_2|) - \left(\frac{|B_1|}{|B_1 \cup B_2|}\right)^\beta \mu(|B_1|) - \left(\frac{|B_2|}{|B_1 \cup B_2|}\right)^\beta \mu(|B_2|).$$

Substituting the expressions above in

$$\mathcal{H}_\beta(\pi') + \left(\frac{|B_1 \cup B_2|}{|S|}\right)^\beta \mathcal{H}_\beta(\{B_1, B_2\})$$

we obtain $\mathcal{H}_\beta(\pi)$. ∎

Theorem 8.33 allows us to extend Axiom (P3):

**Corollary 8.34.** *Let $B_1, \ldots, B_m$ be $m$ nonempty, disjoint sets and let $\pi_i \in PART(B_i)$ for $1 \leq i \leq m$. We have*

$$\mathcal{H}_\beta(\pi_1 + \cdots + \pi_m) = \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i) + \mathcal{H}_\beta(\{B_1, \ldots, B_m\}),$$

*where $S = B_1 \cup \cdots \cup B_m$.*

*Proof.* The argument is by induction on $m \geq 2$. The basis step, $m = 2$, is Axiom (P3). Suppose that the statement holds for $m$, and let $B_1, \ldots, B_m, B_{m+1}$ be $m+1$ disjoint sets. Further, suppose that $\pi_1, \ldots, \pi_m, \pi_{m+1}$ are partitions of these sets, respectively. Then, $\pi_m + \pi_{m+1}$ is a partition of the set $B_m \cup B_{m+1}$. By the inductive hypothesis, we have

$$\begin{aligned}&\mathcal{H}_\beta(\pi_1 + \cdots + (\pi_m + \pi_{m+1})) \\ &= \sum_{i=1}^{m-1} \left(\frac{|B_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i) + \left(\frac{|B_m| + |B_{m+1}|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_m + \pi_{m+1}) \\ &\quad + \mathcal{H}_\beta(\{B_1, \ldots, (B_m \cup B_{m+1})\}),\end{aligned}$$

where $S = B_1 \cup \cdots \cup B_m \cup B_{m+1}$.

Axiom (P3) implies

$$\begin{aligned}
&\mathcal{H}_\beta(\pi_1 + \cdots + (\pi_m + \pi_{m+1})) \\
&= \sum_{i=1}^{m-1} \left(\frac{|B_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i) + \left(\frac{|B_m|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_m) \\
&\quad + \left(\frac{|B_{m+1}|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{m+1}) + \left(\frac{|B_m| + |B_{m+1}|}{|S|}\right)^\beta \mathcal{H}_\beta\{B_m, B_{m+1}\} \\
&\quad + \mathcal{H}_\beta(\{B_1, \ldots, (B_m \cup B_{m+1})\}).
\end{aligned}$$

Finally, an application of Theorem 8.33 gives the desired equality. ∎

**Theorem 8.35.** *Let $\mu$ be the kernel of a $(\Phi, \beta)$-system. If $a, b \in \mathbb{N}_1$, then*

$$\mu(ab) - \mu(a) \cdot b^{1-\beta} = \mu(b).$$

*Proof.* Let $A = \{x_1, \ldots, x_a\}$ and $B = \{y_1, \ldots, y_b\}$ be two nonempty sets. Observe that $\omega_A \times \alpha_B$ consists of $b$ blocks of size $a$: $A \times \{y_1\}, \ldots, A \times \{y_b\}$. By Axiom (P4),

$$\begin{aligned}
&\mathcal{H}_\beta(\omega_A \times \alpha_B) \\
&= \Phi(\mathcal{H}_\beta(\omega_A), \mathcal{H}_\beta(\alpha_B)) = \Phi(0, \mathcal{H}_\beta(\alpha_B)) = \mathcal{H}_\beta(\alpha_B) = \mu(b).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathcal{H}_\beta(\omega_A \times \alpha_B) &= \mathcal{H}_\beta(\alpha_{A\times B}) - \sum_{i=1}^{b} \left(\frac{1}{b}\right)^\beta \mathcal{H}_\beta(\alpha_{A\times\{y_i\}}) \\
&= \mu(ab) - \frac{1}{b^\beta} b \cdot \mu(a),
\end{aligned}$$

which gives the needed equality. ∎

An entropy is said to be *non-Shannon* if it is defined by a $(\Phi, \beta)$-system of axioms such that $\beta > 1$; otherwise (that is if $\beta = 1$), the entropy will be referred to as a *Shannon* entropy. As we shall see, the choice of the parameter $\beta$ determines the form of the function $\Phi$.

Initially we focus on non-Shannon entropies, that is, on $(\Phi, \beta)$-entropies, where $\beta > 1$.

**Theorem 8.36.** *Let $\mathcal{H}_\beta$ be a non-Shannon entropy defined by a $(\Phi, \beta)$-system of axioms and let $\mu$ be the kernel of this system of axioms.*

*There is a number $k > 0$ such that $\mu(a) = k \cdot (1 - a^{1-\beta})$ for every $a \in \mathbb{N}_1$.*

*Proof.* Theorem 8.35 implies that

$$\mu(ab) = \mu(a) \cdot b^{1-\beta} + \mu(b) = \mu(b) \cdot a^{1-\beta} + \mu(a)$$

for every $a, b \in \mathbb{N}_1$. Consequently,

$$\frac{\mu(a)}{1-a^{1-\beta}} = \frac{\mu(b)}{1-b^{1-\beta}} = k$$

for every $a, b \in \mathbb{N}_1$, which gives the desired equality. □

**Corollary 8.37.** *If $\mathcal{H}_\beta$ is a non-Shannon entropy defined by a $(\Phi, \beta)$-system of axioms and $\pi \in PART(S)$, where $\pi = \{B_1, \ldots, B_m\}$, then there exists a constant $k \in \mathbb{R}$ such that*

$$\mathcal{H}_\beta(\pi) = k\left(1 - \sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^\beta\right). \tag{8.2}$$

*Proof.* By Corollary 8.32 and Theorem 8.36, we have

$$\begin{aligned}
\mathcal{H}_\beta(\pi) &= \mu(|S|) - \sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^\beta \mu(|B_i|)\\
&= k\left(1 - \frac{1}{|S|^{\beta-1}}\right) - k\sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^\beta \cdot \left(1 - \frac{1}{|B_i|^{\beta-1}}\right)\\
&= k\left(1 - \frac{1}{|S|^{\beta-1}}\right) - k\sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^\beta + k\sum_{i=1}^{m}\frac{|B_i|}{|S|^\beta}\\
&= k\left(1 - \sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^\beta\right).
\end{aligned}$$

The last equality follows from the fact that $\sum_{i=1}^{m}|B_i| = |S|$. ∎

The constant $k$ introduced in Theorem 8.36 is given by

$$k = \lim_{a\to\infty}\mu(a), \tag{8.3}$$

and the range of values assumed by $\mu$ is $[0, k]$.

Our axiomatization defines entropies (and therefore the kernel $\mu$) up to the multiplicative constant $k$ and the Equality (8.3) expresses this constant in terms of the limit of $\mu(a)$ when $a$ tends to infinity.

The next theorem shows that the function $\Phi$ introduced by Definition 8.28 and used in Axiom (P4) is essentially determined by the choices made for $\beta$ and $k$.

**Theorem 8.38.** *Let $\mathcal{H}_\beta$ be the non-Shannon entropy defined by a $(\Phi, \beta)$-system and let $k$ be as defined by Equality* (8.3)*, where $\mu$ is the kernel of the $(\Phi, \beta)$-system of axioms.*

*The function $\Phi$ of Axiom (P4) is given by $\Phi(x, y) = x + y - \frac{1}{k}\cdot xy$ for $x, y \in \mathbb{R}_{\geq 0}$.*

*Proof.* Let $\pi = \{B_1, \ldots, B_m\} \in PART(S)$ and $\sigma = \{C_1, \ldots, C_n\} \in PART(T)$ be two partitions. Since

$$\sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^{\beta} = 1 - \frac{1}{k}\mathcal{H}_{\beta}(\pi),$$
$$\sum_{j=1}^{n} \left(\frac{|C_j|}{|T|}\right)^{\beta} = 1 - \frac{1}{k}\mathcal{H}_{\beta}(\sigma),$$

we can write

$$\begin{aligned}\mathcal{H}_{\beta}(\pi \times \sigma) &= k\left(1 - \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{|B_i||C_j|}{|S||T|}\right)^{\beta}\right)\\ &= k\left(1 - \left(1 - \frac{1}{k}\mathcal{H}_{\beta}(\pi)\right)\left(1 - \frac{1}{k}\mathcal{H}_{\beta}(\sigma)\right)\right)\\ &= \mathcal{H}_{\beta}(\pi) + \mathcal{H}_{\beta}(\sigma) - \frac{1}{k}\mathcal{H}_{\beta}(\pi)\mathcal{H}_{\beta}(\sigma).\end{aligned}$$

Suppose initially that $\beta > 1$. Observe that the set of rational numbers of the form

$$1 - \sum_{l=1}^{n} r_l^{\beta},$$

where $r_l \in \mathbb{Q}$, $0 \leq r_l \leq 1$ for $1 \leq l \leq n$ and $\sum_{l=1}^{n} r_l = 1$, for some $n \in \mathbb{N}_1$, is dense in the interval $[0, 1]$. Thus, Formula (8.2) shows that the set of entropy values is dense in the interval $[0, k]$ because the sets $B_1, \ldots, B_m$ are finite but of arbitrarily large cardinalities. Since the set of values of entropies is dense in the interval $[0, k]$, the continuity of $\Phi$ implies the desired form of $\Phi$. □

Choosing $k = \frac{1}{1-2^{1-\beta}}$ in Equality (8.2), we obtain the Havrda-Charvat entropy (see [67]):

$$\mathcal{H}_{\beta}(\pi) = \frac{1}{1 - 2^{1-\beta}} \cdot \left(1 - \sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^{\beta}\right).$$

If $\beta = 2$, we obtain $\mathcal{H}_2(\pi)$, which is twice the Gini index,

$$\mathcal{H}_{\beta}(\pi) = 2 \cdot \left(1 - \sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^{2}\right).$$

The *Gini index*, $gini(\pi) = 1 - \sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^2$, is widely used in machine learning and data mining.

The limit case, $\lim_{\beta \to 1} \mathcal{H}_{\beta}(\pi)$, yields

$$\begin{aligned}\lim_{\beta\to 1}\mathcal{H}_\beta(\pi) &= \lim_{\beta\to 1}\frac{1}{1-2^{1-\beta}}\cdot\left(1-\sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^\beta\right)\\ &= \lim_{\beta\to 1}\frac{1}{2^{1-\beta}\ln 2}\cdot\left(-\sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^\beta\ln\frac{|B_i|}{|S|}\right)\\ &= -\sum_{i=1}^{m}\frac{|B_i|}{|S|}\log_2\frac{|B_i|}{|S|},\end{aligned}$$

which is the Shannon entropy of $\pi$.

When $\beta = 1$, by Theorem 8.35, we have

$$\mu(ab) = \mu(a) + \mu(b)$$

for $a, b \in \mathbb{N}_1$. If $\eta : \mathbb{N}_1 \longrightarrow \mathbb{R}$ is the function defined by $\eta(a) = a\mu(a)$ for $a \in \mathbb{N}_1$, then $\eta$ is clearly an increasing function and we have

$$\eta(ab) = ab\mu(ab) = b\eta(a) + a\eta(b)$$

for $a, b \in \mathbb{N}_1$. By Theorem D.6, there exists a constant $c \in \mathbb{R}$ such that $\eta(a) = ca\log_2 a$ for $a \in \mathbb{N}_1$, so $\mu(a) = c\log_2(a)$. Then, Equation (8.1) implies:

$$\mathcal{H}_\beta(\pi) = c\cdot\sum_{i=1}^{m}\frac{a_i}{a}\log_2\frac{a_i}{a}$$

for every partition $\pi = \{A_1, \ldots, A_m\}$ of a set $A$, where $|A_i| = a_i$ for $1 \leq i \leq m$ and $|A| = a$. This is exactly the expression of Shannon's entropy.

The continuous function $\Phi$ is determined as in the previous case. Indeed, if $A, B$ are two sets such that $|A| = a$ and $|B| = b$, then we must have

$$c\cdot\log_2 ab = \mathcal{H}_\beta(\alpha_A\times\alpha_B) = \Phi(c\cdot\log_2 a, c\cdot\log_2 b)$$

for any $a, b \in \mathbb{N}_1$ and any $c \in \mathbb{R}$. The continuity of $\Phi$ implies $\Phi(x, y) = x + y$.

The $\beta$-entropy of $\alpha_S$ is given by

$$\mathcal{H}_\beta(\alpha_S) = \frac{1-|S|^\beta}{1-2^\beta}. \tag{8.4}$$

The entropies previously introduced generate corresponding conditional entropies.

Let $\pi \in PART(S)$ and let $C \subseteq S$. Denote by $\pi_C$ the "trace" of $\pi$ on $C$ given by

$$\pi_C = \{B\cap C | B\in\pi \text{ such that } B\cap C\neq\emptyset\}.$$

Clearly, $\pi_C \in PART(C)$; also, if $C$ is a block of $\pi$, then $\pi_C = \omega_C$.

**Definition 8.39.** *Let $\pi, \sigma \in PART(S)$ and let $\sigma = \{C_1, \ldots, C_n\}$. The* $\beta$-conditional entropy *of the partitions $\pi, \sigma \in PART(S)$ is the function $\mathcal{H}_\beta : PART(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by*

$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_j}).$$

Observe that $\mathcal{H}_\beta(\pi|\omega_S) = \mathcal{H}_\beta(\pi)$ and that $\mathcal{H}_\beta(\omega_S|\pi) = \mathcal{H}_\beta(\pi|\alpha_S) = 0$ for every partition $\pi \in PART(S)$.

For $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$, the conditional entropy can be written explicitly as

$$\begin{aligned}\mathcal{H}_\beta(\pi|\sigma) &= \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \sum_{i=1}^{m} \frac{1}{1-2^{1-\beta}} \left[1 - \left(\frac{|B_i \cap C_j|}{|C_j|}\right)^\beta\right] \\ &= \frac{1}{1-2^{1-\beta}} \sum_{j=1}^{n} \left(\left(\frac{|C_j|}{|S|}\right)^\beta - \sum_{i=1}^{m} \left(\frac{|B_i \cap C_j|}{|S|}\right)^\beta\right). \qquad (8.5)\end{aligned}$$

For the special case when $\pi = \alpha_S$, we can write

$$\mathcal{H}_\beta(\alpha_S|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\alpha_{C_j}) = \frac{1}{1-2^{1-\beta}} \left(\sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta - \frac{1}{|S|^{\beta-1}}\right). \qquad (8.6)$$

**Theorem 8.40.** *Let $\pi$ and $\sigma$ be two partitions of a finite set $S$.*

*We have $\mathcal{H}_\beta(\pi|\sigma) = 0$ if and only if $\sigma \leq \pi$.*

*Proof.* Suppose that $\sigma = \{C_1, \ldots, C_n\}$. If $\sigma \leq \pi$, then $\pi_{C_j} = \omega_{C_j}$ for $1 \leq j \leq n$ and therefore

$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\omega_{C_j}) = 0.$$

Conversely, suppose that

$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_j}) = 0.$$

This implies $\mathcal{H}_\beta(\pi_{C_j}) = 0$ for $1 \leq j \leq n$, which means that $\pi_{C_j} = \omega_{C_j}$ for $1 \leq j \leq n$ by a previous remark. This means that every block $C_j$ of $\sigma$ is included in a block of $\pi$, so $\sigma \leq \pi$. □

The next statement is a generalization of a well-known property of Shannon's entropy.

**Theorem 8.41.** *Let $\pi$ and $\sigma$ be two partitions of a finite set $S$. We have*

$$\mathcal{H}_\beta(\pi \wedge \sigma) = \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma) = \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi),$$

*Proof.* Suppose that $\pi = \{B_1, \ldots, B_m\}$ and that $\sigma = \{C_1, \ldots, C_n\}$. Observe that

$$\pi \wedge \sigma = \pi_{C_1} + \cdots + \pi_{C_n} = \sigma_{B_1} + \cdots + \sigma_{B_m}.$$

Therefore, by Corollary 8.34, we have

$$\mathcal{H}_\beta(\pi \wedge \sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_j}) + \mathcal{H}_\beta(\sigma),$$

which implies

$$\mathcal{H}_\beta(\pi \wedge \sigma) = \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma).$$

The second equality has a similar proof. □

**Corollary 8.42.** *If* $\mathcal{H}_\beta(\pi \wedge \sigma) = \mathcal{H}_\beta(\pi)$, *then* $\pi \leq \sigma$.

*Proof.* Since $\mathcal{H}_\beta(\pi \wedge \sigma) = \mathcal{H}_\beta(\pi)$, Theorem 8.41 implies $\mathcal{H}_\beta(\sigma|\pi) = 0$. By Theorem 8.40, we have $\pi \leq \sigma$. □

**Lemma 8.43.** *Let* $\beta \geq 1$. *If* $w_1, \ldots, w_n$ *are* $n$ *positive numbers such that* $\sum_{k=1}^{n} w_k = 1$ *and* $a_1, \ldots, a_n \in [0, 1]$, *then*

$$1 - \left(\sum_{i=1}^{n} w_i a_i\right)^\beta - \left(\sum_{i=1}^{n} w_i (1 - a_i)\right)^\beta \geq \sum_{i=1}^{n} w_i^\beta \left(1 - a_i^\beta - (1 - a_i)^\beta\right).$$

*Proof.* Let $\phi : [0, 1] \longrightarrow \mathbb{R}$ be the function given by $\phi(x) = x^\beta + (1 - x)^\beta$ for $x \in [0, 1]$. It is easy to see that $\phi(0) = \phi(1) = 1$ and that $\phi$ has a minimum for $x = 1/2$, $\phi(1/2) = 1/2^{1-\beta}$. Thus, we have

$$x^\beta + (1 - x)^\beta \leq 1 \tag{8.7}$$

for $x \in [0, 1]$.

Inequality (8.7) implies

$$w_i(1 - a_i^\beta - (1 - a_i)^\beta) \geq w_i^\beta (1 - a_i^\beta - (1 - a_i)^\beta)$$

because $w_i \in [0, 1]$ and $\beta > 1$.

By applying Jensen's inequality for the convex function $f(x) = x^\beta$, we obtain the inequalities:

$$\left(\sum_{i=1}^{n} w_i a_i\right)^\beta \leq \sum_{i=1}^{n} w_i a_i^\beta,$$

$$\left(\sum_{i=1}^{n} w_i (1 - a_i)\right)^\beta \leq \sum_{i=1}^{n} w_i (1 - a_i)^\beta.$$

Thus, we can write

$$\begin{aligned}
&1-\left(\sum_{i=1}^{n} w_i a_i\right)^{\beta}-\left(\sum_{i=1}^{n} w_i(1-a_i)\right)^{\beta}\\
&=\sum_{i=1}^{n} w_i-\left(\sum_{i=1}^{n} w_i a_i\right)^{\beta}-\left(\sum_{i=1}^{n} w_i(1-a_i)\right)^{\beta}\\
&\geq \sum_{i=1}^{n} w_i-\sum_{i=1}^{n} w_i a_i^{\beta}-\sum_{i=1}^{n} w_i(1-a_i)^{\beta}\\
&=\sum_{i=1}^{n} w_i\left(1-a_i^{\beta}-(1-a_i)^{\beta}\right)\\
&=\sum_{i=1}^{n} w_i^{\beta}\left(1-a_i^{\beta}-(1-a_i)^{\beta}\right),
\end{aligned}$$

which is the desired inequality. □

**Theorem 8.44.** *Let $S$ be a set, $\pi \in PART(S)$ and let $C$ and $D$ be two disjoint subsets of $S$. For $\beta \geq 1$, we have*

$$\left(\frac{|C \cup D|}{|S|}\right)^{\beta} \mathcal{H}_{\beta}(\pi_{C\cup D}) \geq \left(\frac{|C|}{|S|}\right)^{\beta} \mathcal{H}_{\beta}(\pi_C) + \left(\frac{|D|}{|S|}\right)^{\beta} \mathcal{H}_{\beta}(\pi_D).$$

*Proof.* Suppose that $\pi = \{B_1, \ldots, B_m\}$ is a partition of $S$. Define the numbers

$$w_i = \frac{|B_i \cap (C \cup D)|}{|C \cup D|}$$

for $1 \leq i \leq m$. It is clear that $\sum_{i=1}^{m} w_i = 1$. Let

$$a_i = \frac{|B_i \cap C|}{|B_i \cap (C \cup D)|}$$

for $1 \leq i \leq$. It is immediate that $1 - a_i = \frac{|B_i \cap D|}{|B_i \cap (C \cup D)|}$.

Applying Lemma 8.43 to the numbers $w_1, \ldots, w_m$ and $a_1, \ldots, a_m$, we obtain

$$\begin{aligned}
&1-\left(\sum_{i=1}^{n} \frac{|B_i \cap C|}{|C \cup D|}\right)^{\beta}-\left(\sum_{i=1}^{n} \frac{|B_i \cap D|}{|C \cup D|}\right)^{\beta}\\
&\geq \sum_{i=1}^{n} \left(\frac{|B_i \cap (C \cup D)|}{|C \cup D|}\right)^{\beta}\left(1-\left(\frac{|B_i \cap C|}{|B_i \cap (C \cup D)|}\right)^{\beta}-\left(\frac{|B_i \cap D|}{|B_i \cap (C \cup D)|}\right)^{\beta}\right).
\end{aligned}$$

Since

$$\sum_{i=1}^{n} \frac{|B_i \cap C|}{|C \cup D|} = \frac{|C|}{|C \cup D|} \text{ and } \sum_{i=1}^{n} \frac{|B_i \cap D|}{|C \cup D|} = \frac{|D|}{|C \cup D|},$$

the last inequality can be written

$$1-\left(\frac{|C|}{|C\cup D|}\right)^{\beta}-\left(\frac{|D|}{|C\cup D|}\right)^{\beta}$$
$$\geq\sum_{i=1}^{n}\left(\frac{|B_i\cap(C\cup D)|}{|C\cup D|}\right)^{\beta}-\sum_{i=1}^{n}\left(\frac{|B_i\cap C|}{|C\cup D|}\right)^{\beta}-\sum_{i=1}^{n}\left(\frac{|B_i\cap D|}{|C\cup D|}\right)^{\beta},$$

which is equivalent to

$$1-\sum_{i=1}^{n}\left(\frac{|B_i\cap(C\cup D)|}{|C\cup D|}\right)^{\beta}\geq\left(\frac{|C|}{|C\cup D|}\right)^{\beta}\left(1-\sum_{i=1}^{n}\left(\frac{|B_i\cap C|}{|C|}\right)^{\beta}\right)$$
$$+\left(\frac{|D|}{|C\cup D|}\right)^{\beta}\left(1-\sum_{i=1}^{n}\left(\frac{|B_i\cap D|}{|D|}\right)^{\beta}\right),$$

which yields the inequality of the theorem. □

The next result shows that the $\beta$-conditional entropy is dually monotonic with respect to its first argument and is monotonic with respect to its second argument.

**Theorem 8.45.** *Let $\pi,\sigma,\sigma'\in PART(S)$, where $S$ is a finite set. If $\sigma\leq\sigma'$, then $\mathcal{H}_\beta(\sigma|\pi)\geq\mathcal{H}_\beta(\sigma'|\pi)$ and $\mathcal{H}_\beta(\pi|\sigma)\leq\mathcal{H}_\beta(\pi|\sigma')$.*

*Proof.* Since $\sigma\leq\sigma'$, we have $\pi\wedge\sigma\leq\pi\wedge\sigma'$, so $\mathcal{H}_\beta(\pi\wedge\sigma)\geq\mathcal{H}_\beta(\pi\wedge\sigma')$. Therefore, $\mathcal{H}_\beta(\sigma|\pi)+\mathcal{H}_\beta(\pi)\mathcal{H}_\beta(\sigma'|\pi)+\mathcal{H}_\beta(\pi)$, which implies $\mathcal{H}_\beta(\sigma|\pi)\geq\mathcal{H}_\beta(\sigma'|\pi)$.

For the second part of the theorem, it suffices to prove the inequality for partitions $\sigma,\sigma'$ such that $\sigma\prec\sigma'$. Without restricting the generality, we may assume that $\sigma=\{C_1,\ldots,C_{n-2},C_{n-1},C_n\}$ and $\sigma'=\{C_1,\ldots,C_{n-2},C_{n-1}\cup C_n\}$. Thus, we can write

$$\begin{aligned}
&\mathcal{H}_\beta(\pi|\sigma')\\
&=\sum_{j=1}^{n-2}\left(\frac{|C_j|}{|S|}\right)^{\beta}\mathcal{H}_\beta(\pi_{C_j})+\left(\frac{|C_{n-1}\cup C_n|}{|S|}\right)^{\beta}\mathcal{H}_\beta(\pi_{C_{n-1}\cup C_n})\\
&\geq\left(\frac{|C_j|}{|S|}\right)^{\beta}\mathcal{H}_\beta(\pi_{C_j})+\left(\frac{|C_{n-1}|}{|S|}\right)^{\beta}\mathcal{H}_\beta(\pi_{C_{n-1}})+\left(\frac{|C_n|}{|S|}\right)^{\beta}\mathcal{H}_\beta(\pi_{C_n})\\
&\quad\text{(by Theorem 8.44)}\\
&=\mathcal{H}(\pi|\sigma).
\end{aligned}$$

□

**Corollary 8.46.** *We have $\mathcal{H}_\beta(\pi)\geq\mathcal{H}_\beta(\pi|\sigma)$ for every $\pi,\sigma\in PART(S)$.*

*Proof.* We observed that $\mathcal{H}_\beta(\pi)=\mathcal{H}_\beta(\pi|\omega_S)$. Since $\omega_S\geq\sigma$, the statement follows from the second part of Theorem 8.45. □

**Corollary 8.47.** *Let $\xi, \theta, \theta'$ be three partitions of a finite set $S$. If $\theta \geq \theta'$, then*

$$\mathcal{H}_\beta(\xi \wedge \theta) - \mathcal{H}_\beta(\theta) \geq \mathcal{H}_\beta(\xi \wedge \theta') - \mathcal{H}_\beta(\theta').$$

*Proof.* By Theorem 8.41, we have

$$\mathcal{H}_\beta(\xi \wedge \theta) - \mathcal{H}_\beta(\xi \wedge \theta') = \mathcal{H}_\beta(\xi|\theta) + \mathcal{H}_\beta(\theta) - \mathcal{H}_\beta(\xi|\theta') - \mathcal{H}_\beta(\theta').$$

The monotonicity of $\mathcal{H}_\beta(|)$ in its second argument means that: $\mathcal{H}_\beta(\xi|\theta) - \mathcal{H}_\beta(\xi|\theta') \geq 0$, so $\mathcal{H}_\beta(\xi \wedge \theta) - \mathcal{H}_\beta(\xi \wedge \theta') \geq \mathcal{H}_\beta(\theta) - \mathcal{H}_\beta(\theta')$, which implies the desired inequality. □

The behavior of $\beta$-conditional entropies with respect to the "addition" of partitions is discussed in the next statement.

**Theorem 8.48.** *Let $S$ be a finite set and $\pi$ and $\theta$ be two partitions of $S$, where $\theta = \{D_1, \ldots, D_h\}$. If $\sigma_i \in PART(D_i)$ for $1 \leq i \leq h$, then*

$$\mathcal{H}_\beta(\pi|\sigma_1 + \cdots + \sigma_h) = \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{D_i}|\sigma_i).$$

*If $\tau = \{F_1, \ldots, F_k\}$ and $\sigma = \{C_1, \ldots, C_n\}$ are two partitions of $S$, let $\pi_i \in PART(F_i)$ for $1 \leq i \leq k$. Then,*

$$\mathcal{H}_\beta(\pi_1 + \cdots + \pi_k|\sigma) = \sum_{i=1}^{k} \left(\frac{|F_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i|\sigma_{F_i}) + \mathcal{H}_\beta(\tau|\sigma).$$

*Proof.* Suppose that $\sigma_i = \{E_i^\ell \mid 1 \leq \ell \leq p_i\}$. The blocks of the partition $\sigma_1 + \cdots + \sigma_h$ are the sets of the collection $\bigcup_{i=1}^{h} \{E_i^\ell \mid 1 \leq \ell \leq p_i\}$. Thus, we have

$$\mathcal{H}_\beta(\pi|\sigma_1 + \cdots + \sigma_h) = \sum_{i=1}^{h} \sum_{\ell=1}^{p_i} \left(\frac{|E_i^\ell|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{E_i^\ell}).$$

On the other hand, since $(\pi_{D_i})_{E_i^\ell} = \pi_{E_i^\ell}$, we have

$$\begin{aligned}\sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{D_i}|\sigma_i) &= \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta \sum_{\ell=1}^{p_i} \left(\frac{|E_i^\ell|}{|D_i|}\right)^\beta \mathcal{H}_\beta(\pi_{E_i^\ell}) \\ &= \sum_{i=1}^{h} \sum_{\ell=1}^{p_i} \left(\frac{|E_i^\ell|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{E_i^\ell}),\end{aligned}$$

which gives the first equality of the theorem.

To prove the second part, observe that $(\pi_1 + \cdots + \pi_k)_{C_j} = (\pi_1)_{C_j} + \cdots + (\pi_k)_{C_j}$ for every block $C_j$ of $\sigma$. Thus, we have

$$\mathcal{H}_\beta(\pi_1 + \cdots + \pi_k|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta((\pi_1)_{C_j} + \cdots + (\pi_k)_{C_j}).$$

By applying Corollary 8.34 to partitions $(\pi_1)_{C_j}, \ldots, (\pi_k)_{C_j}$ of $C_j$, we can write

$$\mathcal{H}_\beta((\pi_1)_{C_j} + \cdots + (\pi_k)_{C_j}) = \sum_{i=1}^{k} \left( \frac{|F_i \cap C_j|}{|C_j|} \right)^\beta \mathcal{H}_\beta((\pi_i)_{C_j}) + \mathcal{H}_\beta(\tau_{C_j}).$$

Thus,

$$\begin{aligned}
&\mathcal{H}_\beta(\pi_1 + \cdots + \pi_k|\sigma) \\
&= \sum_{j=1}^{n} \sum_{i=1}^{k} \left( \frac{|F_i \cap C_j|}{|S|} \right)^\beta \mathcal{H}_\beta((\pi_i)_{C_j}) + \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^\beta \mathcal{H}_\beta(\tau_{C_j}) \\
&= \sum_{i=1}^{k} \left( \frac{|F_i|}{|S|} \right)^\beta \sum_{j=1}^{n} \left( \frac{|F_i \cap C_j|}{|F_i|} \right)^\beta \mathcal{H}_\beta((\pi_i)_{F_i \cap C_j}) + \mathcal{H}_\beta(\tau|\sigma) \\
&= \sum_{i=1}^{k} \left( \frac{|F_i|}{|S|} \right)^\beta \mathcal{H}_\beta(\pi_i|\sigma_{F_i}) + \mathcal{H}_\beta(\tau|\sigma),
\end{aligned}$$

which is the desired equality. □

**Theorem 8.49.** *Let $\pi, \sigma, \tau$ be three partitions of the finite set $S$. We have*

$$\mathcal{H}_\beta(\pi|\sigma \wedge \tau) + \mathcal{H}_\beta(\sigma|\tau) = \mathcal{H}_\beta(\pi \wedge \sigma|\tau).$$

*Proof.* By Theorem 8.41, we can write

$$\begin{aligned}
\mathcal{H}_\beta(\pi|\sigma \wedge \tau) &= \mathcal{H}_\beta(\pi \wedge \sigma \wedge \tau) - \mathcal{H}_\beta(\sigma \wedge \tau) \\
\mathcal{H}_\beta(\sigma|\tau) &= \mathcal{H}_\beta(\sigma \wedge \tau) - \mathcal{H}_\beta(\tau).
\end{aligned}$$

By adding these equalities and again applying Theorem 8.41, we obtain the equality of the theorem. □

**Corollary 8.50.** *Let $\pi, \sigma, \tau$ be three partitions of the finite set $S$. Then, we have*

$$\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi|\tau).$$

*Proof.* By Theorem 8.49, the monotonicity of $\beta$-conditional entropy in its second argument, and the antimonotonicity of the same in its first argument, we can write

$$\begin{aligned}
\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\tau) &\geq \mathcal{H}_\beta(\pi|\sigma \wedge \tau) + \mathcal{H}_\beta(\sigma|\tau) \\
&= \mathcal{H}_\beta(\pi \wedge \sigma|\tau) \\
&\geq \mathcal{H}_\beta(\pi|\tau),
\end{aligned}$$

which is the desired inequality. □

**Corollary 8.51.** *Let $\pi$ and $\sigma$ be two partitions of the finite set $S$. Then, we have*

$$\mathcal{H}_\beta(\pi \vee \sigma) + \mathcal{H}_\beta(\pi \wedge \sigma) \leq \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma).$$

*Proof.* By Corollary 8.50, we have $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi|\tau) + \mathcal{H}_\beta(\tau|\sigma)$. Then, by Theorem 8.41, we obtain

$$\mathcal{H}_\beta(\pi \wedge \sigma) - \mathcal{H}_\beta(\sigma) \leq \mathcal{H}_\beta(\pi \wedge \tau) - \mathcal{H}_\beta(\tau) + \mathcal{H}_\beta(\tau \wedge \sigma) - \mathcal{H}_\beta(\sigma),$$

hence

$$\mathcal{H}_\beta(\tau) + \mathcal{H}_\beta(\pi \wedge \sigma) \leq \mathcal{H}_\beta(\pi \wedge \tau) + \mathcal{H}_\beta(\tau \wedge \sigma).$$

Choosing $\tau = \pi \vee \sigma$ implies immediately the inequality of the corollary. □

The property of $\mathcal{H}_\beta$ described in Corollary 8.51 is known as the *submodularity* of the generalized entropy. This result generalizes the modularity of the Gini index proven in [118] and gives an elementary proof of a result shown in [89] concerning Shannon's entropy.

## 8.5 Generalized Measures and Data Mining

The notion of a *measure* is important for data mining since, in a certain sense, the support count and the support of an item sets are generalized measures.

The notion of *generalized measure* was introduced in [118], where generalizations of measures of great interest to data mining are considered.

We need first to introduce four properties that apply to real-valued functions defined on a lattice.

**Definition 8.52.** *Let $(L, \{\wedge, \vee\})$ be a lattice. A function $f : L \longrightarrow \mathbb{R}$ is*

- submodular *if $f(u \vee v) + f(u \wedge V) \leq f(u) + f(v)$,*
- supramodular *if $f(u \vee v) + f(u \wedge v) \geq f(u) + f(v)$,*
- logarithmic submodular *if $f(u \vee v) \cdot f(u \wedge v) \leq f(u) \cdot f(v)$, and*
- logarithmic supramodular *if $f(u \vee v) \cdot f(u \wedge v) \geq f(u) \cdot f(v)$,*

*for every $u, v \in L$.*

Clearly, if $f$ is a submodular or supramodular function then $a^f$ is logarithmic submodular or supramodular, respectively, where $a$ is a fixed positive number.

Generalized measures are real-valued functions defined on the lattice of subsets $(\mathcal{P}(S), \{\cup, \cap\})$ of a set $S$. The first two properties introduced in Definition 8.52 may be combined with the monotonicity or antimonotonicity properties to define four types of generalized measures.

**Definition 8.53.** *A* generalized measure *or* g-measure *on a set $S$ is a mapping $m : \mathcal{P}(S) \longrightarrow \mathbb{R}$ that is either monotonic or anti-monotonic and is either submodular or supramodular.*

*Example 8.54.* Let $S$ be a finite nonempty set of nonnegative numbers, $S = \{x_1, x_2, \ldots, x_n\}$ such that $x_1 \leq x_2 \leq \cdots \leq x_n$. Define the mapping max : $\mathcal{P}(S) \longrightarrow \mathbb{R}_{\geq 0}$ by

$$\max(U) = \begin{cases} \text{the largest element of } U & \text{if } U \neq \emptyset, \\ x_1 & \text{if } U = \emptyset, \end{cases}$$

for $U \in \mathcal{P}(S)$.

Note that the definition of max is formulated to ensure that the function is monotonic; that is, $U \subseteq V$ implies $\max U \leq \max V$.

The function max is submodular. Indeed, let $U$ and $V$ be two subsets of $S$ and let $u = \max U$ and $v = \max V$. Without restricting the generality, we may assume that $u \leq v$. In this case, it is clear that $\max(U \cup V) = v$ and that $\max(U \cap V) \leq \max U$ and $\max(U \cap V) \leq \max V$. This implies immediately that max is submodular and therefore that max is a g-measure.

The function min is defined similarly by

$$\min(U) = \begin{cases} \text{the least element of } U & \text{if } U \neq \emptyset, \\ x_n & \text{if } U = \emptyset. \end{cases}$$

The function min is antimonotonic; that is, $U \subseteq V$ implies $\min V \leq \min U$. If $U = \emptyset$, then we have $\emptyset \subseteq V$ for every subset $V$ of $S$ and therefore $\min V \leq \min \emptyset = x_n$, which is obviously true.

It is easy to show that min is a supramodular function, so it is also a g-measure.

Let $f : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geq 0}$ be a nonnegative function defined on the set of subsets of $S$. The functions $f^{\text{neg}}$ and $f^{\text{co}}$ introduced in [118] are defined by

$$f^{\text{neg}}(X) = f(S) + f(\emptyset) - f(X),$$
$$f^{\text{co}}(X) = f(S - X),$$

for $X \in \mathcal{P}(X)$.

**Theorem 8.55.** *Let $f : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geq 0}$ be a nonnegative function defined on the set of subsets of $S$. The following statements are equivalent:*

(i) *$f$ is monotonic.*
(ii) *$f^{neg}$ is antimonotonic.*
(iii) *$f^{co}$ is antimonotonic.*
(iv) *$(f^{co})^{neg}$ is monotonic.*

*Also, the following statements are equivalent:*

(i) *$f$ is submodular.*
(ii) *$f^{neg}$ is supramodular.*
(iii) *$f^{co}$ is submodular.*
(iv) *$(f^{co})^{neg}$ is supramodular.*

*We have* $(f^{neg})^{neg} = f$ *and* $(f^{co})^{co} = f$ *(the* involutive property *of* neg *and* co*).*

*Proof.* The arguments are straightforward and are left to the reader. □

The next result (also from [118]) provides four examples of g-measures.

**Theorem 8.56.** *Let $S$ be a set and let $\mathcal{B}$ be a finite collection of subsets of $S$. Consider the functions $f_{a,supra}, f_{m,supra}, f_{a,sub}, f_{m,sub}$ defined on $\mathcal{P}(S)$:*

$$\begin{aligned} f^{\mathcal{B}}_{a,supra}(X) &= |\{B \in \mathcal{B} | X \subseteq B\}|, \\ f^{\mathcal{B}}_{m,supra}(X) &= |\{B \in \mathcal{B} | \bar{X} \subseteq B\}|, \\ f^{\mathcal{B}}_{a,sub}(X) &= |\{B \in \mathcal{B} | X \not\subseteq B\}|, \\ f^{\mathcal{B}}_{m,sub}(X) &= |\{B \in \mathcal{B} | \bar{X} \not\subseteq B\}|, \end{aligned}$$

*for $X \in \mathcal{P}(S)$. Then, $f^{\mathcal{B}}_{a,supra}$ is antimonotonic and supramodular, $f^{\mathcal{B}}_{m,supra}$ is monotonic and supramodular, $f^{\mathcal{B}}_{a,sub}$ is antimonotonic and submodular, and $f^{\mathcal{B}}_{m,sub}$ is monotonic and submodular, so all four functions are g-measures.*

*Proof.* Since $\{B \in \mathcal{B} | X \cup Y \subseteq B\} \subseteq \{B \in \mathcal{B} | X \subseteq B\}$, it follows that $f^{\mathcal{B}}_{a,supra}(X \cup Y) \leq f^{\mathcal{B}}_{a,supra}(X)$ for every $X, Y \in \mathcal{P}(S)$. Thus, $f^{\mathcal{B}}_{a,supra}(X)$ is antimonotonic.

Observe that

$$\begin{aligned} \{B \in \mathcal{B} | X \cap Y \subseteq B\} &\supseteq \{B \in \mathcal{B} | X \subseteq B\} \cup \{B \in \mathcal{B} | X \subseteq B\}, \\ \{B \in \mathcal{B} | X \cup Y \subseteq B\} &= \{B \in \mathcal{B} | X \subseteq B\} \cap \{B \in \mathcal{B} | X \subseteq B\}. \end{aligned}$$

Therefore,

$$\begin{aligned} & f^{\mathcal{B}}_{a,supra}(X \cap Y) \\ & \geq |\{B \in \mathcal{B} | X \subseteq B\} \cup \{B \in \mathcal{B} | X \subseteq B\}| \\ & = |\{B \in \mathcal{B} | X \subseteq B\}| + |\{B \in \mathcal{B} | X \subseteq B\}| \\ & \quad - |\{B \in \mathcal{B} | X \subseteq B\} \cap \{B \in \mathcal{B} | X \subseteq B\}| \\ & = |\{B \in \mathcal{B} | X \subseteq B\}| + |\{B \in \mathcal{B} | X \subseteq B\}| - |\{B \in \mathcal{B} | X \cup Y \subseteq B\}| \\ & = f^{\mathcal{B}}_{a,supra}(X) + f^{\mathcal{B}}_{a,supra}(Y) - f^{\mathcal{B}}_{a,supra}(X \cup Y), \end{aligned}$$

which yields the supramodular equality.

Observe now that

$$\begin{aligned} f^{\mathcal{B}}_{m,supra} &= (f^{\mathcal{B}}_{a,supra})^{\mathrm{co}}, \\ f^{\mathcal{B}}_{a,sub} &= (f^{\mathcal{B}}_{a,supra})^{\mathrm{neg}}, \\ f^{\mathcal{B}}_{a,supra} &= (f^{\mathcal{B}}_{m,supra})^{\mathrm{neg}}. \end{aligned}$$

The rest of the theorem follows immediately from Theorem 8.55. □

We consider next two important examples of g-measures related to database tables and sets of transactions, respectively. Recall that the partition generated by the set of attributes $X$ of a table was denoted by $\pi^X$.

**Definition 8.57.** *Let $\theta = (T, H, \boldsymbol{r})$ be a table and let $X$ be a set of attributes, $X \subseteq H$. The* $\beta$-entropy of $X$, $\mathsf{H}_\beta(X)$, *is the $\beta$-entropy of the partition of the set of tuples set($\boldsymbol{r}$) generated by $X$:*

$$\mathsf{H}_\beta(X) = \mathcal{H}_\beta(\pi^X).$$

*Example 8.58.* We claim that $\mathsf{H}_\beta$ is a monotonic submodular g-measure on the set of attributes of the table on which it is defined.

Indeed, if $X \subseteq Y$, we saw that $\pi^Y \leq \pi^X$, so $\mathcal{H}_\beta(\pi^X) \leq \mathcal{H}_\beta(\pi^Y)$ by the first axiom of partition entropies. Thus, $\mathsf{H}_\beta$ is monotonic.

To prove the submodularity, we start from the submodularity of the $\beta$-entropy on partitions shown in Corollary 8.51. We have

$$\mathcal{H}_\beta(\pi^X \vee \pi^Y) + \mathcal{H}_\beta(\pi^X \wedge \pi^Y) \leq \mathcal{H}_\beta(\pi^X) + \mathcal{H}_\beta(\pi^Y);$$

hence

$$\mathcal{H}_\beta(\pi^X \vee \pi^Y) + \mathcal{H}_\beta(\pi^{X \cup Y}) \leq \mathcal{H}_\beta(\pi^X) + \mathcal{H}_\beta(\pi^Y)$$

because $\pi^{X \cup Y} = \pi^X \wedge \pi^Y$. Since $X \cap Y \subseteq X$ and $X \cap Y \subseteq Y$ it follows that $\pi^X \leq \pi^{X \cap Y}$ and $\pi^Y \leq \pi^{X \cap Y}$, so $\pi^X \vee \pi^Y \leq \pi^{X \cap Y}$. By Axiom **(P1)**, we have $\mathcal{H}_\beta(\pi^X \vee \pi^Y) \geq \mathcal{H}_\beta(\pi^{X \cap Y})$, which implies

$$\mathcal{H}_\beta(\pi^{X \cap Y}) + \mathcal{H}_\beta(\pi^{X \cup Y}) \leq \mathcal{H}_\beta(\pi^X) + \mathcal{H}_\beta(\pi^Y),$$

which is the submodularity of the g-measure $\mathsf{H}_\beta$.

*Example 8.59.* Let $T$ be a transaction data set over a set of items $I$ as introduced in Definition 7.1. The functions $\mathit{suppcount}_T$ and $\mathit{supp}_T$ introduced in Definition 7.3 are antimonotonic, supramodular g-measures over $\mathcal{P}(I)$. The antimonotonicity of these functions was shown in Theorem 7.6.

Let $K$ and $L$ be two item sets of $T$. If $k$ is the index of a transaction such that either $K \subseteq T(k)$ or $L \subseteq T(k)$, then it is clear that $K \cap L \subseteq T(k)$. Therefore, we have

$$\mathit{suppcount}_T(K \cap L) \geq |\{k \mid K \subseteq T(k)\} \cup \{k \mid L \subseteq T(k)\}|.$$

This allows us to write

$$\begin{aligned}
&\mathit{suppcount}_T(K \cap L)\\
&= |\{k \mid K \cap L \subseteq T(k)\}|\\
&\geq |\{k \mid K \subseteq T(k)\} \cup \{k \mid L \subseteq T(k)\}|\\
&= |\{k \mid K \subseteq T(k)\}| + |\{k \mid L \subseteq T(k)\}|\\
&\quad -|\{k \mid K \subseteq T(k)\} \cap \{k \mid K \subseteq T(k)\}|\\
&= \mathit{suppcount}_T(K) + \mathit{suppcount}_T(L) - \mathit{suppcount}_T(K \cup L)
\end{aligned}$$

for every $K, L \in \mathcal{P}(I)$, which proves that $\mathit{suppcount}_T$ is supramodular. The supramodularity of $\mathit{supp}_T$ follows immediately.

The definition of conditional entropy of partitions allows us to extend this concept to attribute sets of tables.

**Definition 8.60.** *Let $\theta = (T, H, \boldsymbol{r})$ be a table and let $X$ and $Y$ be two attribute sets of this table. The* $\beta$-conditional entropy of $X, Y$ *is defined by*

$$\mathsf{H}_\beta(X|Y) = \mathcal{H}_\beta(\pi^X|\pi^Y).$$

Properties of conditional entropies of partitions can now be easily transferred to conditional entropies of attribute sets. For example, Theorem 8.41 implies

$$\mathsf{H}_\beta(Y|X) = \mathsf{H}_\beta(XY) - \mathsf{H}_\beta(X).$$

## 8.6 Differential Constraints

Differential constraints have been introduced in [119]. They apply to real-valued functions defined over the set of subsets of a set. Examples of such functions are abundant in databases and data mining. For example, we have introduced the entropy of attribute sets mapping sets of attributes into the set of real numbers and the support count of sets of items mapping such sets into natural numbers. Placing restrictions on such functions help us to better express the semantics of data.

**Definition 8.61.** *Let $\mathcal{C}$ be a collection of subsets of a set $S$ and let $f : \mathcal{P}(S) \longrightarrow \mathbb{R}$ be a function. The* $\mathcal{C}$-differential *of $f$ is the function $D_f^{\mathcal{C}} : \mathcal{P}(S) \longrightarrow \mathbb{R}$ defined by*

$$D_f^{\mathcal{C}}(X) = \sum_{\mathcal{D} \subseteq \mathcal{C}} (-1)^{|\mathcal{D}|} f\left(X \cup \bigcup \mathcal{D}\right)$$

*for $X \in \mathcal{P}(S)$.*

*The* density function of $f$ *is the function $d_f : \mathcal{P}(S) \longrightarrow R$ defined by*

$$d_f(X) = \sum_{X \subseteq U \subseteq S} (-1)^{|U|-|X|} f(U) \tag{8.8}$$

*for $X \in \mathcal{P}(S)$.*

Recall that in Example 4.116 we have shown that the Möbius function of a poset $(\mathcal{P}(S), \subseteq)$ is given by

$$\mu(X, U) = \begin{cases} (-1)^{|U|-|X|} & \text{if } X \subseteq U, \\ 0 & \text{otherwise,} \end{cases}$$

for $X, U \in \mathcal{P}(S)$. Therefore, the density $d_f$ can be written as

$$d_f(X) = \sum_{X \subseteq U \subseteq S} \mu(X, U) f(U)$$

for $X, U \in \mathcal{P}(S)$, which implies

$$f(X) = \sum_{X \subseteq U \subseteq S} d_f(U) \tag{8.9}$$

by the Möbius dual inversion theorem (Theorem 4.115).

The density of $f$ can be expressed also as a $\mathcal{C}$-differential. For $X \in \mathcal{P}(S)$, define the collection $\mathcal{C}_X$ as $\mathcal{C}_X = \{\{y\} | y \notin X\}$. Then, we can write

$$\begin{aligned} D_f^{\mathcal{C}_X}(X) &= \sum_{\mathcal{D} \subseteq \mathcal{C}_X} (-1)^{|\mathcal{D}|} f\left(X \cup \bigcup \mathcal{D}\right) \\ &= \sum_{D \subseteq S - X} (-1)^{|D|} f(X \cup D) \\ &= \sum_{X \subseteq U \subseteq S} (-1)^{|U| - |X|} f(U). \end{aligned}$$

In the last equality, we denoted $U = X \cup D$. Since $D$ is a subset of $S - X$, we have immediately $D = U - X$, which justifies the last equality. This allows us to conclude that $d_f(X) = D_f^{\mathcal{C}_X}(X)$ for $X \in \mathcal{P}(S)$.

*Example 8.62.* Let $S$ be a finite set and let $f : \mathcal{P}(S) \longrightarrow \mathbb{R}$. If $\mathcal{C} = \emptyset$, we have $D^{\emptyset}(X) = f(X)$. Similarly, if $\mathcal{C}$ is a one-set collection $\mathcal{C} = \{Y\}$, then $D^{\mathcal{C}}(X) = f(X) - f(X \cup Y)$. When $\mathcal{C} = \{Y, Z\}$, we have $D^{\mathcal{C}}(X) = f(X) - f(X \cup Y) - f(X \cup Z) + f(X \cup Y \cup Z)$ for $X \in \mathcal{P}(S)$.

*Example 8.63.* Let $S = \{a, b, c, d\}$ and let $f : \mathcal{P}(S) \longrightarrow \mathbb{R}$ be a function. For $X = \{a, b\}$, we have $\mathcal{C}_X = \{\{c\}, \{d\}\}$. Thus, $d_f(\{a, b\}) = D_f^{\{c\},\{d\}}(\{a, b\})$. Note that the collections of subsets of $S$ included in the collection $\{\{c\}, \{d\}\}$ are $\emptyset, \{\{c\}\}, \{\{d\}\}$, and $\{\{c\}, \{d\}\}$. Therefore,

$$D_f^{\{\{c\},\{d\}\}}(\{a, b\}) = f(\{a, b\}) - f(\{a, b, c\}) - f(\{a, b, d\}) + f(\{a, b, c, d\}),$$

which equals $d_f(\{a, b\})$, as computed directly from Equality (8.8).

**Definition 8.64.** *Let $\mathcal{C}$ be a collection of subsets of a set $S$. A subset $W$ of $S$ is a* witness set of $\mathcal{C}$ *if $W \subseteq \bigcup \mathcal{C}$ and $X \cap W \neq \emptyset$ for every $X \in \mathcal{C}$.*

*The collection of all witness sets for $\mathcal{C}$ is denoted by $\mathcal{W}(\mathcal{C})$.*

Observe that $\mathcal{W}(\emptyset) = \{\emptyset\}$.

*Example 8.65.* Let $S = \{a, b, c, d\}$ and let $\mathcal{C} = \{\{b\}, \{c, d\}\}$ be a collection of subsets of $S$. The collection of witness sets of $\mathcal{C}$ is

$$\mathcal{W}(\mathcal{C}) = \{\{b, c\}, \{b, d\}, \{b, c, d\}\}.$$

For the collection $\mathcal{D} = \{\{b, c\}, \{b, d\}\}$, we have

$$\mathcal{W}(\mathcal{D}) = \{\{b\}, \{b, c\}, \{b, d\}, \{c, d\}, \{b, c, d\}\}.$$

**Definition 8.66.** *Let $\mathcal{C}$ be a collection of subsets of $S$ and let $X$ be a subset of $S$. The* decomposition *of $\mathcal{C}$ relative to $X$ is the collection $\mathcal{L}[X, \mathcal{C}]$ of subsets of $S$ defined as a union of intervals by*

$$\mathcal{L}[X, \mathcal{C}] = \bigcup_{W \in \mathcal{W}(\mathcal{C})} [X, \overline{W}],$$

*where $\overline{W} = S - W$.*

*Example 8.67.* The decomposition of the collection $\mathcal{C} = \{\{b\}, \{c, d\}\}$ considered in Example 8.65 relative to the set $X = \{a\}$ is given by

$$\begin{aligned}\mathcal{L}[X, \mathcal{C}] &= [\{a\}, \overline{\{b, c\}}] \cup [\{a\}, \overline{\{b, d\}}] \cup [\{a\}, \overline{\{b, c, d\}}] \\ &= [\{a\}, \{a, d\}] \cup [\{a\}, \{a, c\}] \cup [\{a\}, \{a\}] \\ &= \{\{a\}, \{a, d\}, \{a, c\}\}.\end{aligned}$$

Similarly, we can write for the collection $\mathcal{D} = \{\{b, c\}, \{b, d\}\}$

$$\begin{aligned}&\mathcal{L}[X, \mathcal{D}] \\ &= [\{a\}, \overline{\{b\}}] \cup [\{a\}, \overline{\{b, c\}}] \cup [\{a\}, \overline{\{b, d\}}] \cup [\{a\}, \overline{\{c, d\}}] \cup [\{a\}, \overline{\{b, c, d\}}] \\ &= [\{a\}, \{a, c, d\}] \cup [\{a\}, \{a, d\}] \cup [\{a\}, \{a, c\}] \cup [\{a\}, \{a, b\}] \cup [\{a\}, \{a\}] \\ &= \{\{a\}, \{a, c\}, \{a, d\}, \{a, c, d\}, \{a, b\}\}\end{aligned}$$

*Example 8.68.* We have $\mathcal{L}[X, \emptyset] = \bigcup_{W \in \mathcal{W}(\emptyset)} [X, \overline{W}] = [X, S]$ because $\mathcal{W}(\emptyset) = \{\emptyset\}$. Consequently, $\mathcal{L}[\emptyset, \emptyset] = \mathcal{P}(S)$ for every set $S$.

Note that if $X \neq \emptyset$, then $\mathcal{W}(\{X\}) = \mathcal{P}(X) - \{\emptyset\}$. Therefore, $\mathcal{L}[X, \{X\}] = \bigcup_{W \in \mathcal{W}(\mathcal{C})} [X, \overline{W}] = \emptyset$ because there is no set $T$ such that $X \subseteq T \subseteq \overline{W}$.

**Theorem 8.69.** *Let $S$ be a finite set, $X, Y \in \mathcal{P}(S)$, and let $\mathcal{C}$ be a collection of subsets of $S$. We have*

$$\mathcal{L}[X, \mathcal{C}] = \mathcal{L}[X, \mathcal{C} \cup \{Y\}] \cup \mathcal{L}[X \cup Y, \mathcal{C}].$$

*Proof.* We begin the proof by showing that $\mathcal{L}[X, \mathcal{C} \cup \{Y\}] \subseteq \mathcal{L}[X, \mathcal{C}]$. Let $U \in \mathcal{L}[X, \mathcal{C} \cup \{Y\}]$. There is a witness set $W$ for $\mathcal{C} \cup \{Y\}$ such that $X \subseteq U \subseteq \overline{W}$.

The set $W' = W \cap \bigcup \mathcal{C}$ is a witness set for $\mathcal{C}$. Indeed, we have $W' \subseteq \bigcup \mathcal{C}$, and for every set $Z \in \mathcal{C}$ we have $W' \cap Z \neq \emptyset$. Since $W' \subseteq W$, we have $\overline{W} \subseteq \overline{W'}$, so $X \subseteq U \subseteq \overline{W'}$. Therefore, $U \in \mathcal{L}[X, \mathcal{C}]$.

Next, we show that $\mathcal{L}[X \cup Y, \mathcal{C}] \subseteq \mathcal{L}[X, \mathcal{C}]$. Let $V \in \mathcal{L}[X \cup Y, \mathcal{C}]$, so $V \in [X \cup Y, \overline{W}]$ for some witness set of $\mathcal{C}$. Since $[X \cup Y, \overline{W}] \subseteq [X, \overline{W}]$, the desired conclusion follows immediately. Thus, we have shown that

$$\mathcal{L}[X, \mathcal{C} \cup \{Y\}] \cup \mathcal{L}[X \cup Y, \mathcal{C}] \subseteq \mathcal{L}[X, \mathcal{C}].$$

To prove the converse inclusion, let $U \in \mathcal{L}[X, \mathcal{C}]$. There is a witness set $W$ of $\mathcal{C}$ such that $X \subseteq U \subseteq \overline{W}$. Depending on the relative positions of the sets $U$ and $Y$, we can distinguish three cases:

(i) $Y \subseteq U$;
(ii) $Y \not\subseteq U$ and $Y \cap W \neq \emptyset$;
(iii) $Y \not\subseteq U$ and $Y \cap W = \emptyset$.

Note that the first condition of the second case is superfluous because $Y \cap W \neq \emptyset$ implies $Y \not\subseteq U$.

In the first case, we have $U \in \mathcal{L}[X \cup Y, \mathcal{C}]$.

In Case (ii), $W$ is a witness set for $\mathcal{C} \cup \{Y\}$, and therefore $U \in \mathcal{L}[X, \mathcal{C} \cup \{Y\}]$.

Finally, in the third case, define $W_1 = W \cup (Y - U)$. We have $W_1 \subseteq \bigcup \mathcal{C} \cup Y$. Since every member of $\mathcal{C}$ has a nonempty intersection with $W_1$ and $Y \cap W_1 \neq \emptyset$, it follows that $W_1$ is a witness set of $\mathcal{C} \cup \{Y\}$. Note that $U \subseteq \overline{W_1}$. Therefore $U \in \mathcal{L}[X, \mathcal{C} \cup \{Y\}]$. □

The connection between differentials and density functions is shown in the next statement.

**Theorem 8.70.** *Let $S$ be a finite set, $X \in \mathcal{P}(S)$ and let $\mathcal{C}$ be a collection of subsets of $S$. If $f : \mathcal{P}(S) \longrightarrow \mathbb{R}$, then*

$$D_f^{\mathcal{C}}(X) = \sum \{d_f(U) \mid U \in \mathcal{L}[X, \mathcal{C}]\}.$$

*Proof.* By Definition 8.61, the $\mathcal{C}$-*differential* of $f$ is

$$\begin{aligned} D_f^{\mathcal{C}}(X) &= \sum_{\mathcal{D} \subseteq \mathcal{C}} (-1)^{|\mathcal{D}|} f\left(X \cup \bigcup \mathcal{D}\right) \\ &= \sum_{\mathcal{D} \subseteq \mathcal{C}} (-1)^{|\mathcal{D}|} \sum_{X \cup \bigcup \mathcal{D} \subseteq U \subseteq S} d_f(U), \\ &\quad \text{(by Equality 8.9)} \\ &= \sum_{X \subseteq U \subseteq S} d_f(U) \sum_{\mathcal{D} \subseteq \{Y \in \mathcal{C} \mid Y \subseteq U\}} (-1)^{|\mathcal{D}|} \\ &= \sum \{d_f(U) \mid X \subseteq U \subseteq S \text{ and } \{Y \in \mathcal{D} \mid Y \subseteq U\} = \emptyset\}. \end{aligned}$$

□

Constraints can be formulated on real-valued functions defined on collection subsets using differentials of functions or density functions. Both types of constraints have been introduced and studied in [120, 119, 118].

**Definition 8.71.** *Let $S$ be a set, $\mathcal{C}$ a collection of subsets of $S$, and $f : \mathcal{P}(S) \to \mathbb{R}$ a function.*

*The function $f$ satisfies the* differential constraint $X \rightarrowtail \mathcal{C}$ *if $D_f^{\mathcal{C}}(X) = 0$.*

*The function $f$ satisfies the* density constraint $X \rightsquigarrow \mathcal{C}$ *if $d_f(U) = 0$ for every $U \in \mathcal{L}[X, \mathcal{C}]$.*

By Theorem 8.70, if $f$ satisfies the density constraint $X \rightsquigarrow \mathcal{C}$, then $f$ satisfies the differential constraint $X \rightarrowtail \mathcal{C}$. If the density of $f$ takes only nonnegative values (or only nonpositive values), then, by the same theorem, the reverse also holds. However, in general, this is not true, as the next example shows. Thus, differential constraints are weaker than density constraints.

*Example 8.72.* Let $S = \{a\}$ and let $f : \mathcal{P}(S) \to \mathbb{R}$ be defined by $f(\emptyset) = 0$ and $f(\{a\}) = 1$. Observe that $D_f^{\emptyset}(\emptyset) = f(\emptyset) = 0$, so $f$ satisfies the differential constraint $\emptyset \rightarrowtail \emptyset$.

Observe that $\mathcal{C}_{\emptyset} = \{\{a\}\}$ and $\mathcal{C}_{\{a\}} = \emptyset$. Therefore, we have

$$d_f(\emptyset) = D_f^{\mathcal{C}_{\emptyset}}(\emptyset) = f(\emptyset) - f(\{a\}) = -1,$$
$$d_f(\{a\}) = D_f^{\mathcal{C}_{\{a\}}}(\{a\}) = f(\{a\}) = 1.$$

On the other hand, $\mathcal{L}[\emptyset, \emptyset] = \mathcal{P}(S)$ by Example 8.68, and $f$ fails to satisfy the density constraint $\emptyset \rightsquigarrow \emptyset$.

*Example 8.73.* Consider the $\beta$-entropy $\mathsf{H}_\beta$ defined on the set of subsets of the heading of a table $\theta = (T, H, \mathbf{r})$. We saw that $\mathsf{H}_\beta$ is a monotonic, submodular g-measure on $\mathcal{P}(H)$.

We claim that $\mathsf{H}_\beta$ satisfies the differential constraint $X \rightarrowtail \{Y\}$ if and only if the table $\theta$ satisfies the functional dependency $X \to Y$.

By Example 8.62, $\mathsf{H}_\beta$ satisfies the differential constraint $X \rightarrowtail \{Y\}$, that is, $D_f^{\{Y\}}(X) = 0$ if and only if $\mathsf{H}_\beta(X) = \mathsf{H}_\beta(X \cup Y)$. This is equivalent to $\mathcal{H}_\beta(\pi^{XY}) = \mathcal{H}_\beta(\pi^X)$ or to $\mathcal{H}_\beta(\pi^X \wedge \pi^Y) = \mathcal{H}_\beta(\pi^X)$ by Corollary 8.4. This equality implies: $\pi^X \leq \pi^Y$ by Corollary 8.42, which shows that $\theta$ satisfies the functional dependency $X \to Y$.

The observation contained in this example generalizes the result of Sayrafi ([118]) proven for the Gini index and the result contained in [93, 87, 33] that involves the Shannon entropy.

Note also that this shows that

$$\mathsf{H}_\beta(Y|X) = -D_f^{\{Y\}}(X)$$

for $X, Y \in \mathcal{P}(H)$.

*Example 8.74.* Let $S = \{a, b, c\}$. To compute $\mathcal{L}[\{a\}, \{\{b\}\}]$, note that

$$\mathcal{W}(\{\{b\}\}) = \{\{b\}\}.$$

Thus, $\mathcal{L}[\{a\}, \{\{b\}\}] = [\{a\}, \{a, c\}] = \{\{a\}, \{a, c\}\}$. In general, we have for $x, y, z \in S$ that are pairwise distinct

$$\mathcal{L}[\{x\}, \{\{y\}\}] = \{\{x\}, \{x, z\}\}.$$

Consider now a function $f : \mathcal{P}(S) \to \mathbb{R}$ such that

$$f(X) = \begin{cases} 2 & \text{if } X \in \{\emptyset, \{c\}\}, \\ 1 & \text{otherwise.} \end{cases}$$

We have

$$\begin{aligned} d_f(\{c\}) &= f(\{c\}) - f(\{a,c\}) - f(\{b,c\}) + f(\{a,b,c\}) = 1, \\ d_f(\{a,b,c\}) &= f(\{a,b,c\}) = 1. \end{aligned}$$

For $X \notin \{\{c\}, \{a,b,c\}\}$, we have $d_f(X) = 0$. For example,

$$\begin{aligned} d_f(\{b\}) &= f(\{b\}) - f(\{a,b\}) - f(\{b,c\}) + f(\{a,b,c\}) = 0, \\ d_f(\{b,c\}) &= f(\{b,c\}) - f(\{a,b,c\}) = 0. \end{aligned}$$

This shows that $f$ satisfies the density constraints $\{a\} \rightsquigarrow \{\{b\}\}$ and $\{b\} \rightsquigarrow \{\{c\}\}$ but fails to satisfy $\{c\} \rightsquigarrow \{\{a\}\}$ because $\mathfrak{L}[\{c\}, \{\{a\}\}]$ consists of $\{c\}$ and $\{b,c\}$.

## Exercises and Supplements

1. Let $H = A_1 \cdots A_n$ be a set of $n$ attributes. Prove that $|\mathsf{FD}(H)| = 4^n$ and that there exist $4^n - 3^n$ nontrivial functional dependencies in $\mathsf{FD}(H)$.
2. How many functional dependency schemas can be defined on a set $H$ that contains $n$ attributes?
3. Consider the table

$T$

| $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|
| $a_1$ | $b_1$ | $c_1$ | $d_1$ | $e_1$ |
| $a_1$ | $b_1$ | $c_2$ | $d_2$ | $e_2$ |
| $a_1$ | $b_1$ | $c_2$ | $d_3$ | $e_2$ |

   Show several functional dependencies that this table violates.
4. Let $\theta = (T, H, \mathbf{r})$ be a table of a table schema $(H, F)$ such that $\mathbf{r}$ contains no duplicate rows. Prove that the set of attributes $K$ is a key of $\theta$ if and only if $cl_F(K) = H$ and for every proper subset $L$ of $K$ we have $cl_F(L) \subset cl_F(K)$. Formulate and prove a similar statement for reducts.
5. Let $(S, F)$ be a functional dependency schema. Prove that if $A$ is an attribute in $H$ that does not occur on the right member of any functional dependency, then $A$ is a member of the core of the schema.
6. Let $S = (ABCDE, \{AB \to D, BD \to AE, C \to B\})$ be a functional dependency schema. Is the functional dependency $ABC \to DE$ a logical consequence of $F$?
7. Let $S = (A_1 \ldots A_n B_1 \cdots B_n, F)$ be a table schema, where $F = \{A_i \to B_i, B_i \to A_i \mid 1 \leq i \leq n\}$. Prove that any table of this schema has $2^n$ reducts.

8. Let $S$ be a finite set. Prove that for every partition $\pi \in PART(S)$ we have $\mathcal{H}_\beta(\alpha_S|\pi) = \mathcal{H}_\beta(\alpha_S) - \mathcal{H}_\beta(\pi)$.
9. Let $\pi = \{B_1, \ldots, B_m\}$ be a partition of the finite set $S$, where $|S| = n$. Use Jensen's inequality (Theorem B.19) applied to suitable convex functions to prove the following inequalities:
   a) for $\beta > 1$, $\frac{1}{m^{\beta-1}} \leq \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta$;
   b) $\log m \geq -\sum_{i=1}^{m} \frac{|B_i|}{|S|} \log \frac{|B_i|}{|S|}$;
   c) $me^{\frac{1}{m}} \leq \sum_{i=1}^{m} e^{\frac{|B_i|}{|S|}}$.

   **Solution:** Choose, in all cases $p_1 = \cdots = p_m = \frac{1}{m}$. The needed convex functions are $x^\beta$ with $\beta > 1$, $x \log x$, and $e^x$, respectively.
10. Use Supplement 9 to prove that if $\beta > 1$, then $\mathcal{H}_\beta(\pi) \leq \frac{1-m^{1-\beta}}{1-2^{1-\beta}}$.
11. Prove that if $K$ is a reduct of a table $\theta = (T, H, \mathbf{r})$, then $\mathsf{H}_\beta(K) = \min\{\mathsf{H}_\beta(L) \mid L \in \mathcal{P}(H)\}$.

If $\mathcal{L} = (L, \{\wedge, \vee\})$ is a lattice, a mapping $f : L \longrightarrow \mathbb{R}$ is submodular (supramodular) if $f(x \vee y) + f(x \wedge y) \leq f(x) + f(y)$ ($f(x \vee y) + f(x \wedge y) \geq f(x) + f(y)$).

12. Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a lattice and let $f : L \longrightarrow \mathbb{R}$ be an antimonotonic mapping. Prove that the following statements are equivalent:
    a) $f$ is submodular.
    b) $f(z) + f(x \wedge y) \leq f(x \wedge z) + f(z \wedge y)$ for $x, y, z \in L$.

    **Solution:** To prove that (i) implies (ii), apply the submodular inequality to $x \wedge z$ and $z \wedge y$. This yields

$$f((x \wedge z) \vee (z \wedge y)) + f(x \wedge y \wedge z) \leq f(x \wedge z) + f(z \wedge y).$$

    By the subdistributive inequality (5.6) and the anti-monotonicity of $f$, we have

$$f((x \wedge z) \vee (z \wedge y)) \leq f(z \vee (x \wedge y)) \leq f(z).$$

    Since $f(x \wedge y) \leq f(x \wedge y \wedge z)$, the desired inequality follows immediately.

    The reverse implication follows immediately by replacing $z$ by $x \vee y$ and using the absorption properties of the lattice.
13. Let $\mathcal{L} = (L, \{\wedge, \vee\})$ be a lattice and let $f : L \longrightarrow \mathbb{R}$ be an anti-monotonic mapping. Prove that the following statements are equivalent:
    a) $f$ is supramodular.
    b) $f(z) + f(x \vee y) \geq f(x \vee z) + f(z \vee y)$ for $x, y, z \in L$.
14. Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set on the set of items $I$. Prove that if $f = suppcount_T$, then for the density $d_f$ we have $d_f(K) = |\{i \mid T(i) = K\}|$ for every $K \in \mathcal{P}(I)$.
15. Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set on the set of items $I$ and let $\mathcal{C}$ be a collection of sets of items. Prove that $D^{\mathcal{C}}_{suppcount}(K) = \sum\{d_{suppcount}(U) \mid U \in \mathcal{L}[K, \mathcal{C}]\}$.

## Bibliographical Comments

Extensive presentations of functional dependencies and their role in database design are offered in [92, 137, 126].

The identification of functional dependencies satisfied by database tables is a significant topic in data mining [70, 79, 124]. Generalized entropy was introduced in [67] and [34]. The algebraic axiomatization of partition entropy was done in [75] and various applications of Shannon and generalized entropies in data mining were considered in [123, 125].

Generalized measures and their differential constraints were studied in [120, 118, 119].

# 9

# Rough Sets

## 9.1 Introduction

Rough sets are approximative descriptions of sets that can be achieved using equivalences (or partitions). This fertile idea was introduced by the Polish mathematician Z. Pawlak and has generated a large research effort in mathematics and computer science due to its applications.

Unless stated otherwise, all sets in this chapter are finite.

## 9.2 Approximation Spaces

**Definition 9.1.** *Let $S$ be a set. An* approximation space on $S$ *is a pair $(S, \rho)$, where $\rho$ is an equivalence relation defined on the set $S$.*

*If $S$ is clear from the context, we will refer to $(S, \rho)$ just as an approximation space.*

If $(S, \rho)$ is an approximation space defined on $S$ and $U$ is a subset of $S$, then the *$\rho$-degree of membership of an element $x$ of $S$ in $U$* is the number

$$m_\rho(x, U) = \frac{|U \cap [x_\rho]|}{|[x]_\rho|}.$$

Clearly, we have $0 \leq m_\rho(x, U) \leq 1$.

*Example 9.2.* Let $S$ be the set of natural numbers $\{0, 1, \ldots, 12\}$ and let $\rho$ be the equivalence $\equiv_5 \cap (S \times S)$. The equivalence classes of $\rho$ are $\{0, 5, 10\}$, $\{1, 6, 11\}$, $\{2, 7, 12\}$, $\{3, 8\}$, and $\{4, 9\}$.

If $E$ is the subset of even members of $S$, then we have $m_\rho(2, E) = \frac{|E \cap \{2,7,12\}|}{|\{2,7,12\}|} = \frac{2}{3}$ and $m_\rho(4, E) = \frac{|E \cap \{4,9\}|}{|\{4,9\}|} = \frac{1}{2}$.

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_9, © Springer-Verlag London Limited 2008

**Definition 9.3.** *Let $(S, \rho)$ be an approximation space and let $U$ be a subset of $S$. The* $\rho$-lower approximation of $U$ *is the set obtained by taking the union of all $\rho$-equivalence classes included in the set $U$:*

$$lap_\rho(U) = \bigcup\{[x]_\rho \in S/\rho \mid [x]_\rho \subseteq U\}.$$

*The* $\rho$-upper approximation of $U$ *is the set obtained by taking the union of $\rho$-equivalence classes that have a nonempty intersection with the set $U$:*

$$uap_\rho(U) = \bigcup\{[x]_\rho \in S/\rho \mid [x]_\rho \cap U \neq \emptyset\}.$$

*The* $\rho$-boundary of $U$ *is the set*

$$bd_\rho(U) = uap_\rho(U) - lap_\rho(U).$$

If $x \in lap_\rho(U)$, then $x \in U$ and $m_\rho(x, U) = 1$. Thus, $lap_\rho(U)$ is a strong approximation of the set $U$ that consists of those objects of $S$ that can be identified as members of $U$. This set is also known as the *$\rho$-positive region of* $U$ and denoted alternatively by $POS_\rho(U)$.

On the other hand, if $x \in uap_\rho(U)$, then $x$ may or may not belong to $U$. Thus, $uap_\rho(U)$ contains those objects of $S$ that may be members of $U$ and we have $0 \leq m_\rho(x, U) \leq 1$. For $x \in S - uap_\rho(U)$ we have $x \notin U$. This justifies naming the set $S - uap_\rho(U)$ the *$\rho$-negative region of* $U$.

Note that, in general, $lap_\rho(U) \subseteq uap_\rho(U)$ for any set $U$.

The equivalence $\rho$ shall be used interchangeably with the partition $\pi_\rho$ in the notations introduced in Definition 9.3. For example, we can write

$$lap_\rho(U) = \bigcup\{B \in \pi_\rho \mid B \subseteq U\}$$

and

$$uap_\rho(U) = \bigcup\{B \in \pi_\rho \mid B \cap U \neq \emptyset\}.$$

**Definition 9.4.** *Let $(S, \rho)$ be an approximation space. A set $U$, $U \subseteq S$ is* $\rho$-rough *if $bd_\rho(U) \neq \emptyset$ and is* $\rho$-crisp *otherwise.*

*Example 9.5.* Let $S$ be a set and $\rho$ be an equivalence such that the corresponding partition $\pi$ consists of 12 blocks, $B_1, \ldots, B_{12}$ (see Figure 9.1). For the set $U$ shown in this figure, we have

$$\begin{aligned} lap_\rho(U) &= \{B_5, B_{12}\}, \\ uap_\rho(U) &= \{B_1, B_2, B_4, B_5, B_6, B_7, B_8, B_9, B_{10}, B_{12}\}, \\ bd_\rho(U) &= \{B_1, B_2, B_4, B_6, B_7, B_8, B_9, B_{10}\}. \end{aligned}$$

Thus, $U$ is a $\rho$-rough set.

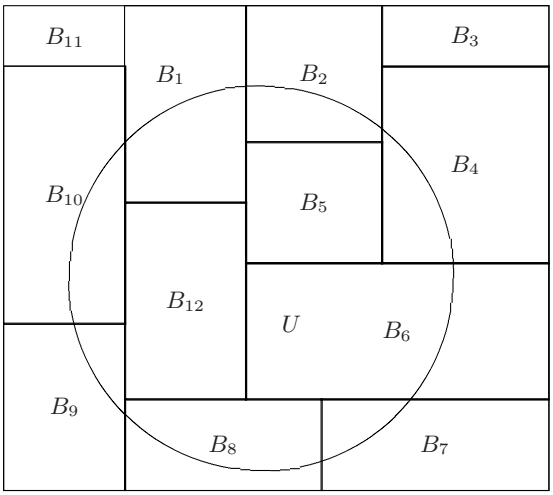



**Fig. 9.1.** Lower and upper approximations of set $U$.

The notion of a $\rho$-saturated set was introduced in Chapter 1. The next statement links this notion to the notion of a $\rho$-crisp set.

**Theorem 9.6.** *Let $(S, \rho)$ be an approximation space. A subset $U$ of $S$ is $\rho$-crisp if and only if $\rho$ is a $\pi_\rho$-saturated set.*

*Proof.* Let $U$ be a $\rho$-crisp set. Since $bd_\rho(U) = uap_\rho(U) - lap_\rho(U) = \emptyset$, it follows that $uap_\rho(U) = lap_\rho(U)$. Thus, $[x]_\rho \cap U \neq \emptyset$ implies $[x]_\rho \subseteq U$. Clearly, if $u \in U$, then $u \in [u]_\rho \cap U$ and therefore $[u]_\rho \subseteq U$, which implies $\bigcup_{u \in U} [u]_\rho \subseteq U$. The reverse inclusion is obvious, so $\bigcup_{u \in U} [u]_\rho = U$, which means that $U$ is $\rho$-saturated.

Conversely, suppose that $U$ is $\rho$-saturated; that is, $\bigcup_{u \in U} [u]_\rho = U$. If $x \in uap_\rho(U)$, then $[x]_\rho \cap U \neq \emptyset$, which means that $[x]_\rho \cap [u]_\rho \neq \emptyset$ for some $u \in U$. Since two equivalence classes that have a nonempty intersection must be equal, it follows that $[x]_\rho = [u]_\rho \subseteq U$, so $x \in lap_\rho(U)$. □

**Theorem 9.7.** *The following statements hold in an approximation space $(S, \rho)$:*

(i) $lap_\rho(\emptyset) = uap_\rho(\emptyset) = \emptyset$ *and* $lap_\rho(S) = uap_\rho(S) = S$,
(ii) $lap_\rho(U \cap V) = lap_\rho(U) \cap lap_\rho(V)$,
(iii) $uap_\rho(U \cup V) = uap_\rho(U) \cup uap_\rho(V)$,
(iv) $lap_\rho(U \cup V) \supseteq lap_\rho(U) \cup lap_\rho(V)$,
(v) $uap_\rho(U \cap V) \subseteq uap_\rho(U) \cap uap_\rho(V)$,
(vi) $lap_\rho(U^c) = \big(uap_\rho(U)\big)^c$ *and* $uap_\rho(U^c) = \big(lap_\rho(U)\big)^c$,

(vii) $lap_\rho(lap_\rho(U)) = uap_\rho(lap_\rho(U)) = lap_\rho(U)$, *and*
(viii) $uap_\rho(uap_\rho(U)) = lap_\rho(uap_\rho(U)) = uap_\rho(U)$
*for every* $U, V \in \mathcal{P}(S)$.

*Proof.* We leave the verification of these statements to the reader. □

**Corollary 9.8.** *Let* $(S, \rho)$ *be an approximation space and let* $U$ *and* $V$ *be two subsets of* $S$. *If* $U \subseteq V$, *then* $lap_\rho(U) \subseteq lap_\rho(V)$ *and* $uap_\rho(U) \subseteq uap_\rho(V)$.

*Proof.* If $U \subseteq V$, we have $U = U \cap V$, so by Part (iii) of Theorem 9.7, we have $lap_\rho(U) = lap_\rho(U) \cap lap_\rho(V)$, which implies $lap_\rho(U) \subseteq lap_\rho(V)$. The second part of this statement follows from Part (iv) of the same theorem. □

**Definition 9.9.** *Let* $(S, \rho)$ *be an approximation space. A subset* $U$ *of* $S$ *is* $\rho$-definable *if* $lap_\rho(U) \neq \emptyset$ *and* $uap_\rho(U) \neq S$.

If $U$ is not $\rho$-definable, then we say that $U$ is *$\rho$-undefinable*. In this case, three cases may occur:

1. If $lap_\rho(U) = \emptyset$ and $uap_\rho(U) \neq S$, then we say that $U$ is *internally $\rho$-undefinable*.
2. If $lap_\rho(U) \neq \emptyset$ and $uap_\rho(U) = S$, then $U$ as an *externally $\rho$-undefinable set*.
3. If $lap_\rho(U) = \emptyset$ and $uap_\rho(U) = S$, then $U$ is a *totally $\rho$-undefinable set*.

**Definition 9.10.** *Let* $(S, \rho)$ *be a finite approximation space on* $S$. *The* accuracy of the $\rho$-approximation *of* $U$ *is the number*

$$\mathsf{acc}\rho(U) = \frac{|lap_\rho(U)|}{|uap_\rho(U)|}.$$

It is clear that $0 \leq \mathsf{acc}\rho(U) \leq 1$. If $\mathsf{acc}\rho(U) = 1$, $U$ is a $\rho$-crisp set; otherwise (that is, if $\mathsf{acc}\rho(U) < 1$), $U$ is *$\rho$-rough*.

*Example 9.11.* Let $S$ be the set of natural numbers $\{0, 1, \ldots, 12\}$ and let $\rho$ be the equivalence $\equiv_5 \cap (S \times S)$ considered in Example 9.2. For the set $E$ of even members of $S$, we have $uap_\rho(E) = \emptyset$ and $lap_\rho(E) = S$, so the set $E$ is a totally $\rho$-undefinable set.

On the other hand, for the subset of perfect squares in $S$, $P = \{1, 4, 9\}$, we have

$$lap_\rho(P) = \{4, 9\},$$
$$uap_\rho(P) = \{1, 4, 9, 6, 11\}.$$

Thus, the accuracy of the $\rho$-approximation of $P$ is $\mathsf{acc}_\rho(P) = 0.4$.

The notion of a $\rho$-positive subset of a set is extended to equivalences as follows.

**Definition 9.12.** *Let $S$ be a set, $\rho$ and $\rho'$ be two equivalences on $S$, and $\pi = \{B_1, \ldots, B_m\}$, $\sigma = \{C_1, \ldots, C_n\}$ the partitions that correspond to $\rho$ and $\rho'$, respectively. The* positive set of $\rho'$ relative to $\rho$ *is the subset of $S$ defined by*

$$POS_\rho(\rho') = \bigcup_{j=1}^{n} lap_\rho(C_j).$$

**Theorem 9.13.** *Let $S$ be a set, and $\rho$ and $\rho'$ two equivalences on $S$. We have $\rho \leq \rho'$ if and only if $POS_\rho(\rho') = S$.*

*Proof.* Let $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$ be the partitions that correspond to $\rho$ and $\rho'$, respectively.

Suppose that $\rho \leq \rho'$. Then, each block $C_j$ of $\sigma$ is a union of blocks of $\pi$, so $lap_\rho(C_j) = C_j$. Therefore, we have

$$POS_\rho(\rho') = \bigcup_{j=1}^{n} lap_\rho(C_j) = \bigcup_{j=1}^{n} C_j = S.$$

Conversely, suppose that $POS_\rho(\rho') = S$, that is, $\bigcup_{j=1}^{n} lap_\rho(C_j) = S$. Since we have $lap_\rho(C_j) \subseteq C_j$ for $1 \leq j \leq n$, we claim that we must have $lap_\rho(C_j) = C_j$ for every $j$, $1 \leq j \leq n$. Indeed, if we have a strict inclusion $lap_\rho(C_{j_0}) \subset C_{j_0}$ for some $j_0$, this implies $\bigcup_{j=1}^{n} lap_\rho(C_j) \subset \bigcup_{j=1}^{n} C_j = S$, which would contradict the equality $\bigcup_{j=1}^{n} lap_\rho(C_j) = S$. Therefore, we have $lap_\rho(C_j) = C_j$ for every $j$, which shows that each block of $\sigma$ is a union of blocks of $\pi$. Consequently, $\rho \leq \rho'$. □

## 9.3 Decision Systems and Decision Trees

Classifiers are algorithms that place objects in certain classes based on characteristic features of those objects. Frequently classifiers are constructed starting from a set of objects known as a *training set*; for each object of the training set the class of the object is known and the classifier can be regarded as a function that maps as well as possible the objects of a training set to their respective classes.

It is desirable that the classifiers constructed from a training set do a reliable job of placing objects that do not belong to the training set in their correct classes. When the classifier works well on the training set but does a poor job of classifying objects outside the training set, we say that the classifier overfits the training set.

Decision systems use tables to formalize the notion of a training set. The features of the objects are represented by the attributes of the table; a special attribute (called a decision attribute) represents the class of the objects.

**Definition 9.14.** *A* decision system *is a pair* $\mathcal{D} = (\theta, D)$*, where* $\theta = (T, H, \boldsymbol{r})$ *and* $D$ *is a special attribute of* $H$ *called a* decision attribute*. The attributes of* $H$ *that are distinct from* $D$ *are referred to as* conditional attributes*, and the set of conditional attributes of* $H$ *will be denoted by* $H_c$*.*

Clearly, $H_c$ is obtained by removing $D$ from $H$.

*Example 9.15.* The decision system considered in this example is based on a data set that is well-known in the machine-learning literature (see [99, 110]). The heading $H$ and domains of the attributes are specified below:

| **Attribute** | **Domain** |
|---|---|
| Outlook | $\{\text{sunny}, \text{overcast}, \text{rain}\}$ |
| Temperature | $\{\text{hot}, \text{mild}, \text{cool}\}$ |
| Humidity | $\{\text{normal}, \text{high}\}$ |
| Wind | $\{\text{weak}, \text{strong}\}$ |

The decision attribute is PlayTennis; this attribute has the domain $\{\text{yes}, \text{no}\}$.

The sequence $\mathbf{r}$ consists of 14 tuples, $t_1, \ldots, t_{14}$, shown in Table 9.1:

**Table 9.1.** The content of the decision system.

| | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 3 | overcast | hot | high | weak | yes |
| 4 | rain | mild | high | weak | yes |
| 5 | rain | cool | normal | weak | yes |
| 6 | rain | cool | normal | strong | no |
| 7 | overcast | cool | normal | strong | yes |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |
| 10 | rain | mild | normal | weak | yes |
| 11 | sunny | mild | normal | strong | yes |
| 12 | overcast | mild | high | strong | yes |
| 13 | rain | hot | normal | weak | yes |
| 14 | rain | mild | high | strong | no |

The partitions of the form $\pi^A$ (where $A$ is an attribute) are

$$\begin{aligned}
\pi^{Outlook} &= \{\{1,2,8,9,11\},\{3,7,12\},\{4,5,6,10,13,14\}\},\\
\pi^{Temperature} &= \{\{1,2,3,13\},\{4,8,10,11,12,14\},\{5,6,7,9\}\},\\
\pi^{Humidity} &= \{\{1,2,3,4,8,12,14\},\{5,6,7,9,10,11,13\}\},\\
\pi^{Wind} &= \{\{1,3,4,5,8,9,10,13\},\{2,6,7,11,12,14\}\},\\
\pi^{PlayTennis} &= \{\{1,2,6,8,14\},\{3,4,5,7,9,10,11,12,13\}\}.
\end{aligned}$$

Let $\mathcal{D} = (\theta, D)$ be a decision system, where $\theta = (T, H, \mathbf{r})$ and $\mathbf{r} = (t_1, \ldots, t_n)$. The *decision function of* $\mathcal{D}$ is the function $\delta_{\mathcal{D}} : \{1, \ldots, n\} \longrightarrow \mathrm{Dom}(D)$ given by

$$\delta_{\mathcal{D}}(i) = \{d \in \mathrm{Dom}\, D \mid (i, j) \in \epsilon^{H_c} \text{ and } t_j[D] = d\},$$

where $\epsilon^{H_c}$ is the indiscernibility relation defined by the set of conditional attributes of $\mathcal{D}$. Equivalently, we can write

$$\delta_{\mathcal{D}}(i) = \{t_j[D] \mid j \in [i]_{\epsilon^{H_c}}\}$$

for $1 \leq i \leq n$.

If $|\delta_{\mathcal{D}}(i)| = 1$ for every $i$, $1 \leq i \leq n$, then $\mathcal{D}$ is a *deterministic (consistent) decision system*; otherwise, $\mathcal{D}$ is a nondeterministic (inconsistent) decision system. In other words, a decision system $\mathcal{D}$ is consistent if the values of the components of a tuple that correspond to the conditional attributes determine uniquely the value of the tuple for the decision attribute.

If there exists $d \in \mathrm{Dom}(D)$ such that $\delta_{\mathcal{D}}(i) = \{d\}$ for every $i$, $1 \leq i \leq n$, then $\mathcal{D}$ is a *pure decision system*.

**Definition 9.16.** *Let $\mathcal{D} = (\theta, D)$ be a decision system, where $\theta = (T, H, \boldsymbol{r})$ and $|\boldsymbol{r}| = n$. The* classification generated by $\mathcal{D}$ *is the partition $\pi^D$ of the set $\{1, \ldots, n\}$.*

*If $B_d$ is the block of $\pi^D$ that corresponds to the value $d$ of* $\mathrm{Dom}(D)$*, we refer to $B_d$ as the $d$-decision class of $\mathcal{D}$.*

Note that the partition $\pi^D$ contains a block for every element of $\mathrm{Dom}(D)$ that occurs in $set(\mathbf{r}[D])$.

If $X$ is a set of attributes, we denote the functions $uap_{\epsilon^X}$ and $lap_{\epsilon^X}$ by $\overline{X}$ and $\underline{X}$, respectively.

*Example 9.17.* The decision classes of the decision system $\mathcal{D}$ of Example 9.15 are

$$B_{no} = \{1, 2, 6, 8, 14\},$$
$$B_{yes} = \{3, 4, 5, 7, 9, 10, 11, 12, 13\}.$$

**Definition 9.18.** *Let $U$ be a set of conditional attributes of $\mathcal{D} = (\theta, D)$, where $\theta = (T, H, \boldsymbol{r})$ and $|\boldsymbol{r}| = n$. The $U$-positive region of the decision system $\mathcal{D}$ is the set*

$$POS_U(\mathcal{D}) = \bigcup \{\underline{U}(B_d) \mid d \in set(\boldsymbol{r}[D])\}.$$

The tuples whose indexes occur in $POS_{H_c}(\mathcal{D})$ can be unambiguously placed in the $d$-decision classes of $\mathcal{D}$.

*Example 9.19.* For the decision system $\mathcal{D}$ of Example 9.15, we have

$$POS_{Outlook}(\mathcal{D}) = \{3, 7, 12\},$$
$$POS_{Temperature}(\mathcal{D}) = POS_{Humidity}(\mathcal{D}) = POS_{Wind}(\mathcal{D}) = \emptyset.$$

Thus, using the value of a single attribute, we can reach a classification decision only for the tuples $t_3, t_7$,and $t_{12}$ (based on the *Outlook* attribute).

Next, we attempt to classify tuples using partitions generated by two attributes. We have six such partitions:

$$\begin{aligned}
\pi^{Outlook,Temperature} &= \{\{1,2\},\{3\},\{4,10,14\},\{5,6\},\\
&\quad \{7\},\{8,11\},\{9\},\{12\},\{13\}\},\\
\pi^{Outlook,Humidity} &= \{\{1,2,8\},\{3,12\},\{4,14\},\{5,6,10,13\},\\
&\quad \{7\},\{9,11\}\},\\
\pi^{Outlook,Wind} &= \{\{1,8,9\},\{2,11\},\{3\},\\
&\quad \{4,5,10,13\},\{6,14\},\{7,12\}\},\\
\pi^{Temperature,Humidity} &= \{\{1,2,3\},\{4,8,12,14\},\{5,6,7,9\},\\
&\quad \{10,11\},\{13\}\},\\
\pi^{Temperature,Wind} &= \{\{1,3,13\},\{2\},\{4,8,10\},\\
&\quad \{5,9\},\{6,7\},\{11,12,14\}\},\\
\pi^{Humidity,Wind} &= \{\{1,3,4,8\},\{2,12,14\},\{5,9,10,13\},\\
&\quad \{6,7,11\}\},
\end{aligned}$$

and their corresponding positive regions are

$$\begin{aligned}
POS_{Outlook,Temperature}(\mathcal{D}) &= \{1,2,3,7,9,13\},\\
POS_{Outlook,Humidity}(\mathcal{D}) &= \{1,2,3,7,8,9,11,12\},\\
POS_{Outlook,Wind}(\mathcal{D}) &= \{3,4,5,6,7,10,12,13,14\},\\
POS_{Temperature,Humidity}(\mathcal{D}) &= \{10,11,13\},\\
POS_{Temperature,Wind}(\mathcal{D}) &= \{2,5,9\},\\
POS_{Humidity,Wind}(\mathcal{D}) &= \{5,9,10,13\}.
\end{aligned}$$

There are four partitions generated by three attributes:

$$\begin{aligned}
\pi^{Outlook,Temperature,Humidity} &= \{\{1,2\},\{3\},\{4,14\},\{5,6\},\\
&\quad \{7\},\{8\},\{9\},\{10\},\{11\},\{12\},\{13\}\},\\
\pi^{Outlook,Temperature,Wind} &= \{\{1\},\{2\},\{3\},\{4,10\},\{5\},\{6\},\\
&\quad \{7\},\{8\},\{9\},\{11\},\{12\},\{13\},\{14\}\},\\
\pi^{Outlook,Humidity,Wind} &= \{\{1,8\},\{2\},\{3\},\{4\},\{5,10,13\},\\
&\quad \{6\},\{7\},\{9\},\{11\},\{12\},\{14\}\},\\
\pi^{Temperature,Humidity,Wind} &= \{\{1,3\},\{2\},\{4,8\},\{5,9\},\\
&\quad \{6,7\},\{10\},\{11\},\{12,14\},\{13\}\}.
\end{aligned}$$

Their positive regions are

$$\begin{aligned} POS_{Outlook,Temperature,Humidity}(\mathcal{D}) &= \{1,2,3,7,8,9,10,11,12,13\},\\ POS_{Outlook,Temperature,Wind}(\mathcal{D}) &= set(H),\\ POS_{Outlook,Humidity,Wind}(\mathcal{D}) &= set(H),\\ POS_{Temperature,Humidity,Wind}(\mathcal{D}) &= \{2,5,9,10,11,13\}. \end{aligned}$$

This computation shows that a classification decision can be reached for every tuple starting from its components on the projection on either the set Outlook, Temperature, Wind or the set Outlook, Humidity, Wind.

Finally, note that $\mathcal{D}$ is a deterministic system because $\pi^{set(H_c)} = \alpha_{set(H_c)}$.

**Theorem 9.20.** *Let $\mathcal{D} = (\theta, D)$ be a decision system, where $\theta = (T, H, \boldsymbol{r})$ and $|\boldsymbol{r}| = n$. The following statements are equivalent:*
(i) *$\mathcal{D}$ is deterministic;*
(ii) *$\pi^{H_c} \leq \pi^D$;*
(iii) *$POS_{H_c}(\mathcal{D}) = \{1, \ldots, n\}$.*

*Proof.* (i) implies (ii): Suppose that $\mathcal{D}$ is deterministic. Then, if two tuples $u$ and $v$ in $\mathbf{r}$ are such that $u[H_c] = v[H_c]$, then $u[D] = v[D]$. This is equivalent to saying that $\pi^{H_c} \leq \pi^D$.

(ii) implies (iii): This implication follows immediately from Theorem 9.13.

(iii) implies (i): Suppose that $POS_{H_c}(\mathcal{D}) = \{1, \ldots, n\}$, that is,

$$\bigcup \{\underline{H_c}(B_d) \mid d \in set(\mathbf{r}[D])\} = \{1, \ldots, n\}. \tag{9.1}$$

Note that $\underline{H_c}(B_d) = lap_{\epsilon^{H_c}}(B_d) \subseteq B_d$. Therefore, we have $lap_{\epsilon^{H_c}}(B_d) = B_d$ for every $d \in set(\mathbf{r}[D])$ because if the inclusion were strict for any of the sets $B_d$, then Equality 9.1 would be violated. Thus, each block $B_d$ is a union of blocks of the partition $\pi^{H_c}$. In other words, each equivalence class of $\epsilon^{H_c}$ is included in a block $B_d$, which means that for every $j \in [i]_{\epsilon^{H_c}}$ we have $t_j[D] = d$ and the set $\delta_{\mathcal{D}}(i)$ contains a single element. Thus, $\mathcal{D}$ is deterministic. □

Next, we discuss informally a classification algorithm that makes use of decision systems. This algorithm is recursive and begins with the selection of a conditional attribute (referred to as a *splitting attribute*) by applying a criterion that is specific to the algorithm. The algorithm halts when it applies to a pure decision system (or to a decision system where a minimum percentage of tuples have the same decision value).

The splitting attribute is chosen here as one of the attributes whose positive region has the largest number of elements. This choice is known as a *splitting criterion*. The table of the decision system is split into a number of tables such that each new table is characterized by a value of the splitting attribute. Thus, we obtain a new set of decision systems, and the algorithm is applied recursively to the new decision systems. The process must stop because the tables become smaller with each split; its result is a tree of decision systems known as a *decision tree*.

*Example 9.21.* As we saw in Example 9.19, a classification decision can be reached immediately for the tuples $t_3, t_7, t_{12}$, which are characterized by the condition Outlook = 'overcast'.

Define the tables

$$\theta_0 = (\theta \textbf{ where } \text{Outlook} = \text{'sunny'})[K],$$
$$\theta_1 = (\theta \textbf{ where } \text{Outlook} = \text{'overcast'})[K],$$
$$\theta_2 = (\theta \textbf{ where } \text{Outlook} = \text{'rain'})[K],$$

where

$$K = \text{ Temperature Humidity Wind PlayTennis}$$

is the heading obtained by dropping the Outlook attribute and the decision systems $\mathcal{D}_i = (\theta_i, D)$ for $0 \leq i \leq 3$.

Note that the decision system $\mathcal{D}_1$ is pure because the tuples of $\theta_1$ belong to the same PlayTennis-class as shown in Table 9.2.

**Table 9.2.** The table $\theta_1$.

| | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|
| 3 | hot | high | weak | yes |
| 7 | cool | normal | strong | yes |
| 12 | mild | high | strong | yes |

For the remaining systems

$$\mathcal{D}_0 = (\theta_0, \text{PlayTennis}) \text{ and } \mathcal{D}_2 = (\theta_2, \text{PlayTennis}),$$

we have the tables shown in Tables 9.3(a) and (b).

The same process is now applied to the decision systems $\mathcal{D}_0$ and $\mathcal{D}_2$. The positive regions are:

$$POS_{Temperature}(\mathcal{D}_0) = \{1, 2, 9\},$$
$$POS_{Humidity}(\mathcal{D}_0) = \{1, 2, 8, 9, 11\},$$
$$POS_{Wind}(\mathcal{D}_0) = \emptyset,$$
$$POS_{Temperature}(\mathcal{D}_2) = \{13\},$$
$$POS_{Humidity}(\mathcal{D}_2) = \emptyset,$$
$$POS_{Wind}(\mathcal{D}_2) = \{4, 5, 10, 13, 6, 14\}.$$

Thus, the splitting attribute for $\mathcal{D}_0$ is *Humidity*; the splitting attribute for $\mathcal{D}_2$ is *Wind*.

The decision system $\mathcal{D}_0$ yields the decision systems

$$\mathcal{D}_{00} = (\theta_{00}, PlayTennis) \text{ and } \mathcal{D}_{01} = (\theta_{01}, PlayTennis),$$

where $\theta_{00}$ and $\theta_{01}$ are given by

**Table 9.3.** The tables $\theta_0$ and $\theta_2$.

| | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|
| 1 | hot | high | weak | no |
| 2 | hot | high | strong | no |
| 8 | mild | high | weak | no |
| 9 | cool | normal | weak | yes |
| 11 | mild | normal | strong | yes |

(a)

| | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|
| 4 | mild | high | weak | yes |
| 5 | cool | normal | weak | yes |
| 6 | cool | normal | strong | no |
| 10 | mild | normal | weak | yes |
| 13 | hot | normal | weak | yes |
| 14 | mild | high | strong | no |

(b)

$$\theta_{00} = (\theta_0 \textbf{ where } \text{Humidity} = \text{'high'})[K_0],$$
$$\theta_{01} = (\theta_0 \textbf{ where } \text{Humidity} = \text{'normal'})[K_0],$$

where

$$K_0 = \text{ Temperature Wind PlayTennis.}$$

The tables $\theta_{00}$ and $\theta_{01}$ are shown in Tables 9.4(a) and (b), respectively. Note that both decision systems $\mathcal{D}_{00}$ and $\mathcal{D}_{01}$ are pure, so no further action is needed.

**Table 9.4.** The tables $\theta_{00}$ and $\theta_{01}$.

| | Temperature | Wind | PlayTennis |
|---|---|---|---|
| 1 | hot | weak | no |
| 2 | hot | strong | no |
| 8 | mild | weak | no |

(a)

| | Temperature | Wind | PlayTennis |
|---|---|---|---|
| 9 | cool | weak | yes |
| 11 | mild | strong | yes |

(b)

The decision system $\mathcal{D}_2$ produces the decision systems

$$\mathcal{D}_{20} = (\theta_{20}, PlayTennis) \text{ and } \mathcal{D}_{21} = (\theta_{21}, PlayTennis),$$

where $\theta_{20}$ and $\theta_{21}$ are given by

$$\theta_{20} = (\theta_0 \textbf{ where } \text{Wind} = \text{'weak'})[K_2]$$
$$\theta_{21} = (\theta_0 \textbf{ where } \text{Wind} = \text{'strong'})[K_2],$$

where

$$K_2 = \text{ Temperature Humidity PlayTennis.}$$

The tables $\theta_{20}$ and $\theta_{21}$ are shown in Tables 9.5(a) and (b), respectively.

**Table 9.5.** The tables $\theta_{00}$ and $\theta_{01}$.

| | Temperature | Humidity | PlayTennis |
|---|---|---|---|
| 4 | mild | high | yes |
| 5 | cool | normal | yes |
| 10 | mild | normal | yes |
| 13 | hot | normal | yes |

(a)

| | Temperature | Humidity | PlayTennis |
|---|---|---|---|
| 6 | cool | normal | no |
| 14 | mild | high | no |

(b)

Again, both decision systems are pure, so no further splitting is necessary. The decision tree that is obtained from this process is shown in Figure 9.2.

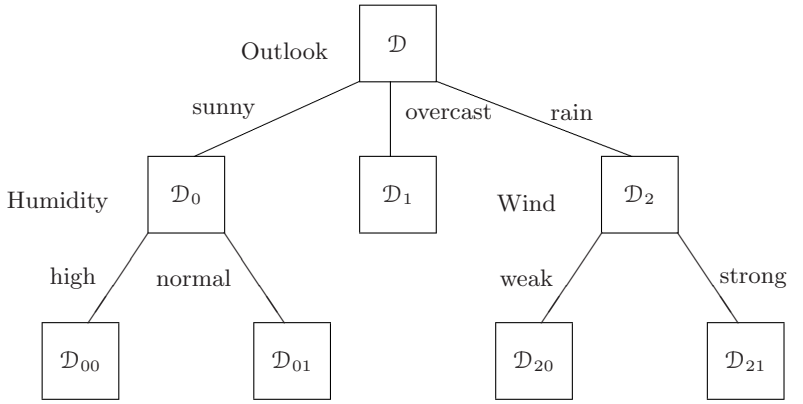



**Fig. 9.2.** Decision tree.

## 9.4 Closure Operators and Rough Sets

In Exercise 10, the reader is asked to prove that the lower approximation operator $lap_\rho$ defined by an equivalence on a set $S$ is an interior operator on $S$ and the upper approximation operator $uap_\rho$ is a closure operator on the same set. In addition, it is easy to see (Exercise 2) that

$$\begin{aligned} uap_\rho(\emptyset) &= \emptyset, \\ uap_\rho(U \cup V) &= uap_\rho(U) \cup uap_\rho(V), \\ U &\subseteq S - uap_\rho(S - uap_\rho(U)), \\ lap_\rho(S) &= S, \\ lap_\rho(U \cap V) &= lap_\rho(U) \cap lap_\rho(V), \\ U &\supseteq S - lap_\rho(S - lap_\rho(U)), \end{aligned}$$

for every $U, V \in \mathcal{P}(S)$.

It has been shown (see [146]) that approximation spaces can be defined starting from certain closure operators or interior operators.

**Theorem 9.22.** *Let $S$ be a set and let $\boldsymbol{K}$ be a closure operator on $S$ such that the following conditions are satisfied:*
(i) $\boldsymbol{K}(\emptyset) = \emptyset$,
(ii) $\boldsymbol{K}(U \cup V) = \boldsymbol{K}(U) \cup \boldsymbol{K}(V)$, *and*
(iii) $U \subseteq S - \boldsymbol{K}(S - \boldsymbol{K}(U))$,
*for every $U, V \in \mathcal{P}(S)$. Then, the mapping $\boldsymbol{I} : \mathcal{P}(S) \longrightarrow \mathcal{P}(S)$ defined by $\boldsymbol{I}(U) = S - \boldsymbol{K}(S - U)$ for $U \in \mathcal{P}(S)$ is an interior operator on $S$ and there exists an approximation space $(S, \rho)$ such that $lap_\rho(U) = \boldsymbol{I}(U)$ and $uap_\rho(U) = \boldsymbol{K}(U)$ for every $U \in \mathcal{P}(S)$.*

*Proof.* Define the relation $\rho$ by

$$\rho = \{(x, y) \in S \times S \mid x \in \mathbf{K}(\{y\})\}.$$

Denote the set $\{v \in S \mid u \in \mathbf{K}(\{v\})\}$ by $r(u)$ for $u \in S$.

Observe that we have

$$\mathbf{K}(\{y\}) = \{x \in S \mid y \in r(x)\}$$

for $x, y \in S$. We begin the argument by proving that $\rho$ is an equivalence.

The reflexivity of $\rho$ follows from $x \in \mathbf{K}(\{x\})$ for $x \in S$.

We claim that $y \notin \mathbf{K}(W)$ if and only if $r(y) \cap W = \emptyset$. Since $\mathbf{K}(W) = \bigcup\{\mathbf{K}(\{x\}) \mid x \in W\}$, it follows that the statements
(i) $y \notin \mathbf{K}(W)$,
(ii) $y \notin \mathbf{K}(\{x\})$ for every $x \in W$,
(iii) $x \notin r(y)$ for every $x \in W$, and
(iv) $r(y) \cap W = \emptyset$

are equivalent, which justifies our claim.

Property (iii) of Theorem 9.22 implies that $\{y\} \subseteq S - \mathbf{K}(S - \mathbf{K}(\{y\}))$; that is, $y \notin \mathbf{K}(S - \mathbf{K}(\{y\}))$. Therefore, by the argument of the previous paragraph, we have $r(y) \cap (S - \mathbf{K}(\{y\})) = \emptyset$, which implies $r(y) \subseteq \mathbf{K}(\{y\})$. Thus, if $x \in r(y)$, it follows that $x \in \mathbf{K}(\{y\})$; that is, $y \in r(x)$. In terms of the relation $\rho$, this means that $(y, x) \in \rho$ implies $(x, y) \in \rho$, so $\rho$ is a symmetric relation.

To prove the transitivity of $\rho$, suppose that $(x, y), (y, z) \in \rho$. We have $x \in \mathbf{K}(\{y\})$ and $y \in \mathbf{K}(\{z\})$. By the idempotency of $\mathbf{K}$, we have $\mathbf{K}(\{y\}) \subseteq \mathbf{K}(\mathbf{K}(\{z\})) = \mathbf{K}(\{z\})$, so $x \in \mathbf{K}(\{z\})$. Consequently, $(x, z) \in \rho$. This allows us to conclude that $\rho$ is indeed an equivalence. Moreover, we realize now that $r(x)$ is exacly the equivalence class $[x]_\rho$ for $x \in S$.

An immediate consequence of this fact is that we have $x \in r(y)$ if and only if $y \in r(x)$. Therefore,

$$\mathbf{K}(\{y\}) = \{x \in S \mid y \in r(x)\} = \{x \in S \mid x \in r(y)\} = r(y),$$

and by the second property of $\mathbf{K}$, we have

$$\mathbf{K}(U) = \bigcup \{r(u) \mid u \in U\}$$

for $U \in \mathcal{P}(S)$. Consequently, we can write

$$\begin{aligned} uap_\rho(U) &= \bigcup \{r(x) \mid r(x) \cap U \neq \emptyset\} \\ &= \bigcup \{r(u) \mid u \in U\} \\ &= \mathbf{K}(U). \end{aligned}$$

Also, we have $\mathbf{K}(\{z\}) = r(z)$ for every $z \in S$.

The definition of $\mathbf{I}$ implies immediately that this function is an interior operator on $S$ that enjoys two additional properties, namely $\mathbf{I}(S) = S$ and $\mathbf{I}(U \cap V) = \mathbf{I}(U) \cap \mathbf{I}(V)$.

By applying the definition of $\mathbf{I}$, we have

$$\begin{aligned} \mathbf{I}(U) &= S - \mathbf{K}(S - U) \\ &= S - \bigcup \{\mathbf{K}(\{z\}) \mid z \in S - U\} \\ &= S - \bigcup \{r(z) \mid z \in S - U\} \\ &= \bigcup \{r(z) \mid z \in U\} \\ &= lap_\rho(U). \end{aligned}$$

□

## Exercises and Supplements

1. Let $\rho_1$ and $\rho_2$ be two equivalences on a set $S$ such that $\rho_1 \subseteq \rho_2$. Prove that $lap_{\rho_1}(U) \supseteq lap_{\rho_2}(U)$ and $lap_{\rho_1}(U) \subseteq lap_{\rho_2}(U)$. Conclude that $bd_{\rho_1}(U) \subseteq bd_{\rho_2}(U)$ for every $U \in \mathcal{P}(S)$.
2. Let $(S, \rho)$ be an approximation space. Prove that
   a) $uap_\rho(\emptyset) = \emptyset$,
   b) $uap_\rho(U \cup V) = uap_\rho(U) \cup uap_\rho(V)$,
   c) $U \subseteq S - uap_\rho(S - uap_\rho(U))$,
   d) $lap_\rho(S) = S$,
   e) $lap_\rho(U \cap V) = lap_\rho(U) \cap lap_\rho(V)$, and
   f) $U \supseteq S - lap_\rho(S - lap_\rho(U))$

   for every $U, V \in \mathcal{P}(S)$.
3. Let $(S, \rho)$ be an approximation space. A *lower (upper) sample* of a subset $U$ of $S$ is a subset $Y$ of $S$ such that $Y \subseteq U$ and $uap_\rho(Y) = lap_\rho(U)$ ($uap_\rho(Y) = uap_\rho(U)$, respectively). A lower (upper) sample of $U$ is *minimal* if there no lower (upper) sample of $U$ with fewer elements.
   Prove that every nonempty lower (upper) sample $Y$ of a set $U$ has a nonempty intersection with each $\rho$-equivalence class included in $lap_\rho(U)$ ($lap_\rho(U)$, respectively). Prove that if $Y$ is a lower (upper) minimal sample, then its intersection with each $\rho$-equivalence class included in $lap_\rho(U)$ ($lap_\rho(U)$, respectively) consists of exactly one element.

A *generalized approximation space* is a pair $(S, \rho)$, where $\rho$ is an arbitrary relation on $S$. Denote the set $\{y \in S \mid (x, y) \in \rho\}$ by $\rho(x)$. The lower and upper $\rho$-approximations of a set $U \in \mathcal{P}(S)$ generalize the corresponding notions from approximation spaces and are defined by

$$lap_\rho(U) = \bigcup\{\rho(x) \mid \rho(x) \subseteq U\},$$
$$uap_\rho(U) = \bigcup\{\rho(x) \mid \rho(x) \cap U \neq \emptyset\},$$

for $U \in \mathcal{P}(S)$.

4. Let $(S, \rho)$ be a generalized approximation space. Prove that
   a) $lap_\rho(U) = S - uap_\rho(S - U)$,
   b) $lap_\rho(S) = S$,
   c) $lap_\rho(U \cap V) = lap_\rho(U) \cap lap_\rho(V)$,
   d) $lap_\rho(U \cup V) \supseteq lap_\rho(U) \cup lap_\rho(V)$,
   e) $U \subseteq V$ implies $lap_\rho(U) \subseteq lap_\rho(V)$,
   f) $uap_\rho(U) = S - lap_\rho(S - U)$,
   g) $uap_\rho(\emptyset) = \emptyset$,
   h) $uap_\rho(U \cap V) \subseteq uap_\rho(U) \cap uap_\rho(V)$,
   i) $uap_\rho(U \cup V) = uap_\rho(U) \cup uap_\rho(V)$,
   j) $U \subseteq V$ implies $uap_\rho(U) \subseteq uap_\rho(V)$,
   k) $lap_\rho((S - U) \cup V) \subseteq (S - lap_\rho(U)) \cup lap_\rho(V)$

for $U, V \in \mathcal{P}(S)$.

5. Let $(S, \rho)$ be a generalized approximation space, where $\rho$ is a tolerance relation. Prove that
   a) $lap_\rho(\emptyset) = \emptyset$,
   b) $lap_\rho(U) \subseteq U$,
   c) $U \subseteq lap_\rho(uap_\rho(U))$,
   d) $uap_\rho(S) = S$,
   e) $U \subseteq uap_\rho(U)$, and
   f) $uap_\rho(lap_\rho(U)) \subseteq U$

   for $U, V \in \mathcal{P}(S)$.
6. Does every table have a reduct?
7. What can be said about a table $\theta$ that has a nonempty core that is a reduct?
8. Let $\mathcal{D} = (\theta, D)$ be a decision system. A *D-reduct* is a set of attributes $L$, $L \subseteq H_c$ such that $\epsilon^L = \epsilon^D$ and $L$ is minimal with this property. Prove that every reduct of the table $\theta$ includes a $D$-reduct.
9. Let $\mathcal{D} = (\theta, D)$ be a decision system. If $U$ and $V$ are two sets of conditional attributes such that $U \subseteq V$, then prove that $POS_U(\mathcal{D}) \supseteq POS_V(\mathcal{D})$.
10. Let $S$ be a set and let $\rho$ be an equivalence on $S$.
    a) Prove that $lap_\rho$ is an interior operator on $S$.
    b) Prove that $uap_\rho$ is a closure operator on $S$.
    c) Prove that the $lap_\rho$-open subsets of $S$ coincide with the $uap_\rho$-closed subsets of $S$.
11. Let $(S, \mathcal{O})$ be a finite topological space. Prove that there exists a bijection between the set $EQS(S)$ of equivalences on $S$ and the set of topologies on $S$ in which every open set is also closed.

## Bibliographical Comments

Rough sets were introduced by Z. Pawlak (see [106]). Excellent surveys supplemented by large bibliographies are [81] and [42]. The notions of lower and upper samples discussed in Exercise 3 are introduced in [19]. Various generalizations of the notion of approximation space are presented in [146].

# Part III

# Metric Spaces

# 10

# Dissimilarities, Metrics, and Ultrametrics

## 10.1 Introduction

The notion of a metric was introduced in mathematics by the French mathematician Maurice René Fréchet in [53] as an abstraction of the notion of distance between two points. In this chapter, we explore the notion of metric and the related notion of metric space, as well as a number of generalizations and specializations of these notions.

Clustering and classification, two central data mining activities, require the evaluation of degrees of dissimilarity between data objects. This task is accomplished using the notion of dissimilarity and a variety of specializations of this notion, such as metrics, tree metrics, and ultrametrics. These notions are introduced in Section 10.2.

Substantial attention is paid to ultrametrics due to their importance for clustering algorithms. Various modalities for generating ultrametrics are discussed, starting with hierarchies on sets, equidistant trees, and chains of partitions (or equivalences).

Metrics on several quite distinct data types are discussed: vectors in $\mathbb{R}^n$, subsets of finite sets, partitions of finite sets, and sequences. The chapter concludes with a section dedicated to the application of elementary properties of metrics to searching in metric spaces. Further applications of metrics are presented in subsequent chapters.

## 10.2 Classes of Dissimilarities

Dissimilarities are functions that allow us to evaluate the extent to which data objects are different.

**Definition 10.1.** *A* dissimilarity on a set $S$ *is a function* $d : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ *satisfying the following conditions:*
*(DISS1)* $d(x, x) = 0$ *for all* $x \in S$*;*

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_10, © Springer-Verlag London Limited 2008

*(DISS2)* $d(x, y) = d(y, x)$ *for all* $x, y \in S$.

*The pair* $(S, d)$ *is a* dissimilarity space. *A trivial example of a dissimilarity space is the pair* $(\emptyset, d_\emptyset)$, *where* $d_\emptyset$ *is the empty function on* $\emptyset \times \emptyset$.

The set of dissimilarities defined on a set $S$ is denoted by $\mathcal{D}_S$.

Additional properties may be satisfied by dissimilarities. A nonexhaustive list is given next.

1. $d(x, y) = 0$ implies $d(x, z) = d(y, z)$ for every $x, y, z \in S$ (*evenness*).
2. $d(x, y) = 0$ implies $x = y$ for every $x, y$ (*definiteness*).
3. $d(x, y) \leq d(x, z) + d(z, y)$ for every $x, y, z$ (*triangular inequality*).
4. $d(x, y) \leq \max\{d(x, z), d(z, y)\}$ for every $x, y, z$ (*the ultrametric inequality*).
5. $d(x, y)+d(u, v) \leq \max\{d(x, u)+d(y, v), d(x, v)+d(y, u)\}$ for every $x, y, u, v$ (*Buneman's inequality*, also known as the *four-point condition*).

If $d : S^2 \longrightarrow \mathbb{R}$ is a function that satisfies the properties (DISS1), (DISS2), and the triangular inequality, then the values of $d$ are nonnegative numbers. Indeed, by taking $x = y$ in the triangular inequality, we have

$$0 = d(x, x) \leq d(x, z) + d(z, x) = 2d(x, z),$$

for every $z \in S$.

Later in this chapter, we explore various connections that exist among these properties. As an example, we can show the following statement.

**Theorem 10.2.** *Both the triangular inequality and definiteness imply evenness.*

*Proof.* Suppose that $d$ is a dissimilarity that satisfies the triangular inequality, and let $x, y \in S$ be such that $d(x, y) = 0$. By the triangular inequality, we have both $d(x, z) \leq d(x, y) + d(y, z) = d(y, z)$ and $d(y, z) \leq d(y, x) + d(x, z) = d(x, z)$ because $d(y, x) = d(x, y) = 0$. Thus, $d(x, z) = d(y, z)$ for every $z \in S$.

We leave it to the reader to prove the second part of the statement. □

We denote the set of definite dissimilarities on a set $S$ by $\mathcal{D}'_S$. Further notations will be introduced shortly for other types of dissimilarities.

**Definition 10.3.** *A dissimilarity* $d \in \mathcal{D}_S$ *is*

1. *a* metric *if it satisfies the definiteness property and the triangular inequality,*
2. *a* tree metric *if it satisfies the definiteness property and Buneman's inequality, and*
3. *an* ultrametric *if it satisfies the definiteness property and the ultrametric inequality.*

*The set of metrics on a set* $S$ *is denoted by* $\mathcal{M}_S$. *The sets of tree metrics and ultrametrics on a set* $S$ *are denoted by* $\mathcal{T}_S$ *and* $\mathcal{U}_S$, *respectively.*

*If* $d$ *is a metric or an ultrametric on a set* $S$, *then* $(S, d)$ *is a* metric space *or an* ultrametric space, *respectively.*

The pair $(\emptyset, d_\emptyset)$ mentioned before is a trivial example of metric space.

If $d$ is a metric defined on a set $S$ and $x, y \in S$, we refer to the number $d(x, y)$ as the *d-distance* between $x$ and $y$ or simply the *distance* between $x$ and $y$ whenever $d$ is clearly understood from context.

Thus, a function $d : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ is a metric if it has the following properties:

(M1) $d(x, y) = 0$ if and only if $x = y$ for $x, y \in S$;
(M2) $d(x, y) = d(y, x)$ for $x, y \in S$;
(M3) $d(x, y) \leq d(x, z) + d(z, y)$ for $x, y, z \in S$.

If property (M1) is replaced by the weaker requirement that $d(x, x) = 0$ for $x \in S$, then we refer to $d$ as a *semimetric* on $S$. Thus, if $d$ is a semimetric $d(x, y) = 0$ does not necessarily imply $x = y$ and we can have for two distinct elements $x, y$ of $S$, $d(x, y) = 0$.

*Example 10.4.* Let $S$ be a nonempty set. Define the mapping $d : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ by

$$d(u, v) = \begin{cases} 1 & \text{if } u \neq v, \\ 0 & \text{otherwise,} \end{cases}$$

for $x, y \in S$. It is easy to see that $d$ satisfies the definiteness property. To prove that $d$ satisfies the triangular inequality, we need to show that

$$d(x, y) \leq d(x, z) + d(z, y)$$

for all $x, y, z \in S$. This is clearly the case if $x = y$. Suppose that $x \neq y$, so $d(x, y) = 1$. Then, for every $z \in S$, we have at least one of the inequalities $x \neq z$ or $z \neq y$, so at least one of the numbers $d(x, z)$ or $d(z, y)$ equals 1. Thus $d$ satisfies the triangular inequality. The metric $d$ introduced here is the *discrete metric* on $S$.

*Example 10.5.* Consider the mapping $d : (\mathbf{Seq}_n(S))^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by

$$d(\mathbf{p}, \mathbf{q}) = |\{i \mid 0 \leq i \leq n-1 \text{ and } \mathbf{p}(i) \neq \mathbf{q}(i)\}|$$

for all sequences $\mathbf{p}, \mathbf{q}$ of length $n$ on the set $S$.

Clearly, $d$ is a dissimilarity that is both even and definite. Moreover, it satisfies the triangular inequality. Indeed, let $\mathbf{p}, \mathbf{q}, \mathbf{r}$ be three sequences of length $n$ on the set $S$. If $\mathbf{p}(i) \neq \mathbf{q}(i)$, then $\mathbf{r}(i)$ must be distinct from at least one of $\mathbf{p}(i)$ and $\mathbf{q}(i)$. Therefore,

$$\begin{aligned} &\{i \mid 0 \leq i \leq n-1 \text{ and } \mathbf{p}(i) \neq \mathbf{q}(i)\} \\ &\subseteq \{i \mid 0 \leq i \leq n-1 \text{ and } \mathbf{p}(i) \neq \mathbf{r}(i)\} \cup \{i \mid 0 \leq i \leq n-1 \text{ and } \mathbf{r}(i) \neq \mathbf{q}(i)\}, \end{aligned}$$

which implies the triangular inequality.

A function $d : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ is an ultrametric if it has the following properties:

(U1) $d(x, y) = 0$ if and only if $x = y$ for $x, y \in S$;
(U2) $d(x, y) = d(y, x)$ for $x, y \in S$;
(U3) $d(x, y) \leq \max\{d(x, z), d(z, y)\}$ for $x, y, z \in S$.

As we did for metrics, if property (U1) is replaced by the weaker requirement that $d(x, x) = 0$ for $x \in S$, then $d$ is a *quasi-ultrametric* on $S$.

*Example 10.6.* Let $\pi = \{B, C\}$ be a two-set partition of a nonempty set $S$. Define the mapping $d : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ by

$$d(x, y) = \begin{cases} 0 & \text{if } \{x, y\} \subseteq B \text{ or } \{x, y\} \subseteq C \\ 1 & \text{otherwise,} \end{cases}$$

for $x, y \in S$.

We claim that $d$ is a quasi-ultrametric. Indeed, it is clear that $d(x, x) = 0$ for every $x \in S$ and $d(x, y) = d(y, x)$ for $x, y \in S$. Now let $x, y, z$ be three arbitrary elements in $S$. If $d(x, y) = 1$, then $x$ and $y$ belong to two distinct blocks of the partition $\pi$, say to $B$ and $C$, respectively. If $z \in B$, then $d(x, z) = 0$ and $d(z, y) = 1$; similarly, if $z \in C$, then $d(x, z) = 1$ and $d(z, y) = 0$. In either case, the ultrametric inequality is satisfied and we conclude that $d$ is a quasi-ultrametric.

**Theorem 10.7.** . *Let $a_0, a_1, a_2 \in \mathbb{R}$ be three numbers. If $a_i \leq \max\{a_j, a_k\}$ for every permutation $(i, j, k)$ of the set $\{0, 1, 2\}$, then two of the numbers are equal and the third is not larger than the two others.*

*Proof.* Suppose that $a_i$ is the least of the numbers $a_0, a_1, a_2$ and $a_j, a_k$ are the remaining numbers. Since $a_j \leq \max\{a_i, a_k\} = a_k$ and $a_k \leq \max\{a_i, a_j\} = a_j$, it follows that $a_j = a_k \geq a_i$. □

A simple and interesting property of triangles in ultrametric spaces is given next.

**Corollary 10.8.** *Let $(S, d)$ be an ultrametric space. For every $x, y, z \in S$, two of the numbers $d(x, y), d(x, z), d(y, z)$ are equal and the third is not larger than the two other equal numbers.*

*Proof.* Since $d$ satisfies the ultrametric inequality, the statement follows immediately from Theorem 10.7. □

Corollary 10.8 can be paraphrased by saying that in an ultrametric space any triangle is isosceles and the side that is not equal to the two others cannot be longer than these.

In this chapter, we frequently use the notion of a sphere.

**Definition 10.9.** *Let $(S, d)$ be a metric space. The* closed sphere *centered in $x \in S$ of radius $r$ is the set*

$$B_d(x, r) = \{y \in S | d(x, y) \leq r\}.$$

*The* open sphere *centered in $x \in S$ of radius $r$ is the set*

$$C_d(x, r) = \{y \in S | d(x, y) < r\}.$$

**Definition 10.10.** *Let $(S, d)$ be a metric space. The* diameter *of a subset $U$ of $S$ is the number $diam_{S,d}(U) = \sup\{d(x, y) \mid x, y \in U\}$. The set $U$ is* bounded *if $diam_{S,d}(U)$ is finite.*

*The* diameter *of the metric space $(S, d)$ is the number*

$$diam_{S,d} = \sup\{d(x, y) \mid x, y \in S\}.$$

*If the metric space is clear from the context, then we denote the diameter of a subset $U$ just by $diam(U)$.*

*If $(S, d)$ is a finite metric space, then $diam_{S,d} = \max\{d(x, y) \mid x, y \in S\}$.*

A dissimilarity $d : S \times S \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ can be extended to the set of subsets of $S$ by defining $d(U, V)$ as

$$d(U, V) = \inf\{d(u, v) \mid u \in U \text{and } v \in V\}$$

for $U, V \in \mathcal{P}(S)$. The resulting extension is also a dissimilarity. However, even if $d$ is a metric, then its extension is not, in general, a metric on $\mathcal{P}(S)$ because it does not satisfy the triangular inequality. Instead, we can show that for every $U, V, W$ we have

$$d(U, W) \leq d(U, V) + diam(V) + d(V, W).$$

Indeed, by the definition of $d(U, V)$ and $d(V, W)$, for every $\epsilon > 0$, there exist $u \in U$, $v, v' \in V$, and $w \in W$ such that

$$\begin{array}{l} d(U, V) \leq d(u, v) \leq d(U, V) + \frac{\epsilon}{2}, \\ d(V, W) \leq d(v', w) \leq d(V, W) + \frac{\epsilon}{2}. \end{array}$$

By the triangular axiom, we have

$$d(u, w) \leq d(u, v) + d(v, v') + d(v', w).$$

Hence,

$$d(u, w) \leq d(U, V) + diam(V) + d(V, W) + \epsilon,$$

which implies

$$d(U, W) \leq d(U, V) + diam(V) + d(V, W) + \epsilon$$

for every $\epsilon > 0$. This yields the needed inequality.

**Definition 10.11.** *Let $(S, d)$ be a metric space. The sets $U, V \in \mathcal{P}(S)$ are* separate *if $d(U, V) > 0$.*

We denote the number $d(\{u\}, V) = \inf\{d(u, v) \mid v \in V\}$ by $d(u, V)$. It is clear that $u \in V$ implies $d(u, V) = 0$. Further properties of this functions are discussed in Theorem 11.16 on page 427.

Let $d$ be a dissimilarity and let $S(x, y)$ be the set of all nonnull sequences $\mathbf{s} = (s_1, \ldots, s_n) \in \mathbf{Seq}(S)$ such that $s_1 = x$ and $s_n = y$. The *d-amplitude of* $\boldsymbol{s}$ is the number $amp_d(\mathbf{s}) = \max\{d(s_i, s_{i+1}) \mid 1 \leq i \leq n-1\}$.

If $d$ is an ultrametric, we saw that $d(x, y) \leq amp_d(\mathbf{s})$ for any nonnull sequence $\mathbf{s} = (s_1, \ldots, s_n)$ such that $s_1 = x$ and $s_n = y$. Therefore, we have

$$d(x, y) \leq \min\{amp_d(\mathbf{s}) \mid \mathbf{s} \in S(x, y)\},$$

where $S(x, y)$ is the set of sequences of $S$ that start with $x$ and end with $y$. Since $(x, y) \in S(x, y)$, we have the equality

$$d(x, y) = \min\{amp_d(\mathbf{s}) \mid \mathbf{s} \in S(x, y)\}.$$

Dissimilarities defined on finite sets can be represented by matrices. If $S = \{x_1, \ldots, x_n\}$ is a finite set and $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ is a dissimilarity, let $M_d \in (\mathbb{R}_{\geq 0})^{n \times n}$ be the matrix defined by $M_{ij} = d(x_i, x_j)$ for $1 \leq i, j \leq n$. Clearly, all main diagonal elements of $M_d$ are 0 and the matrix $M$ is symmetric.

*Example 10.12.* Let $S$ be the set $\{x_1, x_2, x_3, x_4\}$. The discrete metric on $S$ is represented by the $4 \times 4$-matrix

$$M_d = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

If $x_1, x_2, x_3 \in \mathbb{R}$ are three real numbers the matrix that represents the distance $e(x_i, x_j) = |x_i - x_j|$ measured on the real line is

$$M_e = \begin{pmatrix} 0 & |x_1 - x_2| & |x_1 - x_3| \\ |x_1 - x_2| & 0 & |x_2 - x_3| \\ |x_1 - x_3| & |x_2 - x_3| & 0 \end{pmatrix}.$$

Next we introduce the notion of extended dissimilarity by allowing $\infty$ as a value of a dissimilarity.

**Definition 10.13.** *Let $S$ be a set. An* extended dissimilarity *on $S$ is a function $d : S^2 \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ that satisfies the conditions (DISS1) and (DISS2) of Definition 10.1.*

*The pair $(S, d)$ is an* extended dissimilarity space.

The notions of extended metric and extended ultrametric are defined starting from the notion of extended dissimilarity using the same process as in the definitions of metrics and ultrametrics.

## 10.3 Tree Metrics

The distance $d$ between two vertices of a connected graph $\mathcal{G} = (V, E)$ introduced in Definition 3.12 is a metric on the set of vertices $V$. Recall that $d(x, y) = m$ if $m$ is the length of the shortest path that connects $x$ and $y$.

We have $d(x, y) = 0$ if and only if $x = y$. The symmetry of $d$ is obvious. If $\mathbf{p}$ is a shortest path that connects $x$ to $z$ and $\mathbf{q}$ is a shortest path that connects $z$ to $y$, then $\mathbf{pq}$ is a path of length $d(x, z) + d(z, y)$ that connects $x$ to $y$. Therefore, $d(x, y) \leq d(x, z) + d(z, y)$.

The notion of distance between the vertices of a connected graph can be generalized to weighted graphs as follows. Let $(\mathcal{G}, w)$ be a weighted graph where $\mathcal{G} = (V, E)$, and let $w : E \longrightarrow \mathbb{R}_{\geq 0}$ be a positive weight. Define $d_w(x, y)$ as

$$d_w(x, y) = \min\{w(\mathbf{p}) \mid \mathbf{p} \text{ is a path joining } x \text{ to } y\}$$

for $x, y \in V$. We leave to reader to prove that $d_w$ satisfies the conditions (M1) - (M3) using an argument that is similar to the one given above.

If $(\mathcal{T}, w)$ is a weighted tree the metric $d_w$ is referred to as a *tree metric*. Since $\mathcal{T}$ is a tree, for any two vertices $u, v \in V$ there is a unique simple path $\mathbf{p} = (v_0, \ldots, v_n)$ joining $u = v_0$ to $v = v_n$. In this case,

$$d_w(u, v) = \sum_{i=0}^{n-1} w(v_i, v_{i+1}).$$

Moreover, if $t = v_k$ is a vertex on the path $\mathbf{p}$, then

$$d_w(u, t) + d_w(t, v) = d_w(u, v), \tag{10.1}$$

a property known as the *additivity* of $d_w$.

We already know that $d_w$ is a metric for arbitrary connected graphs. For trees, we have the additional property given in the next statement.

**Theorem 10.14.** *If $(\mathcal{T}, w)$ is a weighted tree, then $d_w$ satisfies Buneman's inequality*

$$d_w(x, y) + d_w(u, v) \leq \max\{d_w(x, u) + d_w(y, v), d_w(x, v) + d_w(y, u)\}$$

*for every four vertices $x, y, u, v$ of the tree $\mathcal{T}$.*

*Proof.* Let $x, y, u, v$ be four vertices in $\mathcal{T}$. If $x = u$ and $y = v$, the inequality reduces to an obvious equality. Therefore, we may assume that at least one of the pairs $(x, u)$ and $(y, v)$ consists of distinct vertices.

Suppose that $x = u$. In this case, the inequality amounts to

$$d_w(x, y) + d_w(x, v) \leq \max\{d_w(y, v), d_w(x, v) + d_w(y, x)\},$$

which is obviously satisfied. Thus, we may assume that we have both $x \neq u$ and $y \neq v$. Since $\mathcal{T}$ is a tree, there exists a simple path $\mathbf{p}$ that joins $x$ to $y$ and a simple path $\mathbf{q}$ that joins $u$ to $v$.

Two cases may occur, depending on whether **p** and **q** have common edges.

Suppose initially that there are no common vertices between **p** and **q**. Let $s$ be a vertex on the path **p** and $t$ be a vertex on **q** such that $d_w(s,t)$ is the minimal distance between a vertex located on the path **p** and one located on **q**; here $d_w(s,t)$ is the sum of the weights of the edges of the simple path **r** that joins $s$ to $t$.

The path **r** has no other vertices in common with **p** and **q** except $s$ and $t$, respectively (see Figure 10.1). We have

$$\begin{aligned}
d_w(x,u) &= d_w(x,s) + d_w(s,t) + d_w(t,u),\\
d_w(y,v) &= d_w(y,s) + d_w(s,t) + d_w(t,v),\\
d_w(x,v) &= d_w(x,s) + d_w(s,t) + d_w(t,v),\\
d_w(y,u) &= d_w(y,s) + d_w(s,t) + d_w(t,u).
\end{aligned}$$

Thus, $d_w(x,u)+d_w(y,v) = d_w(x,v)+d_w(y,u) = d_w(x,s)+d_w(s,t)+d_w(t,u)+ d_w(y,s) + d_w(s,t) + d_w(t,v) = d_w(x,y) + d_w(u,v) + 2d_w(s,t)$, which shows that Buneman's inequality is satisfied.

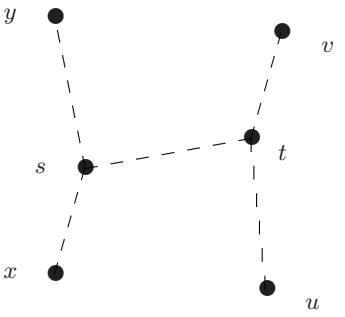



**Fig. 10.1.** Paths that have no common vertices.

If **p** and **q** have some vertices in common, the configuration of the graph is as shown in Figure 10.2. In this case, we have

$$\begin{aligned}
d_w(x,y) &= d_w(x,t) + d_w(t,s) + d_w(s,y),\\
d_w(u,v) &= d_w(u,t) + d_w(t,s) + d_w(s,v),\\
d_w(x,u) &= d_w(x,t) + d_w(t,u),\\
d_w(y,v) &= d_w(y,s) + d_w(s,v),\\
d_w(x,v) &= d_w(x,t) + d_w(t,s) + d_w(s,v),\\
d_w(y,u) &= d_w(y,s) + d_w(s,t) + d_w(t,u),
\end{aligned}$$

which implies

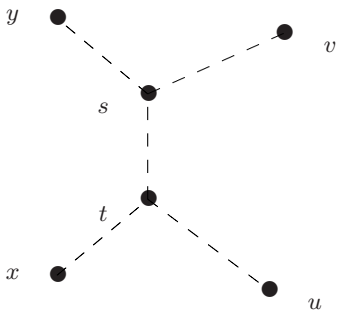



**Fig. 10.2.** Paths that share vertices.

$$d_w(x,y) + d_w(u,v) = d_w(x,t) + 2d_w(t,s) + d_w(s,y) + d_w(u,t) + d_w(s,v),$$
$$d_w(x,u) + d_w(y,v) = d_w(x,t) + d_w(t,u) + d_w(y,s) + d_w(s,v),$$
$$d_w(x,v) + d_w(y,u) = d_w(x,t) + 2d_w(t,s) + d_w(s,v) + d_w(y,s) + d_w(t,u).$$

Thus, Buneman's inequality is satisfied in this case, too, because

$$d_w(x,y) + d_w(u,v) = d_w(x,v) + d(y,u) \geq d_w(x,u) + d_w(y,v).$$

□

By Theorem 10.7, Buneman's inequality is equivalent to saying that of the three sums $d(x,y) + d(u,v)$, $d(x,u) + d(y,v)$, and $d(x,v) + d(y,u)$, two are equal and the third is no less than the two others.

Next, we examine the relationships that exist between metrics, tree metrics, and ultrametrics.

**Theorem 10.15.** *Every tree metric is a metric, and every ultrametric is a tree metric.*

*Proof.* Let $S$ be a nonempty set and let $d$ be a tree metric on $S$, that is, a dissimilarity that satisfies the inequality $d(x,y) + d(u,v) \leq \max\{d(x,u) + d(y,v), d(x,v) + d(y,u)\}$ for every $x, y, u, v \in S$. Choosing $v = u$, we obtain $d(x,y) \leq \max\{d(x,u) + d(y,u), d(x,u) + d(y,u)\} = d(x,u) + d(u,y)$ for every $x, y, u$, which shows that $d$ satisfies the triangular inequality.

Suppose now that $d$ is an ultrametric. We need to show that

$$d(x,y) + d(u,v) \leq \max\{d(x,u) + d(y,v), d(x,v) + d(y,v)\}$$

for $x, y, u, v \in S$. Several cases are possible depending on which of the elements $u, v$ is the closest to $x$ and $y$, as the next table shows:

| Case | Closest to | | Implications |
|---|---|---|---|
| | $x$ | $y$ | |
| 1 | $u$ | $u$ | $d(x,v) = d(u,v)$, $d(y,v) = d(u,v)$ |
| 2 | $u$ | $v$ | $d(x,v) = d(u,v)$, $d(y,u) = d(u,v)$ |
| 3 | $v$ | $u$ | $d(x,u) = d(u,v)$, $d(y,v) = d(u,v)$ |
| 4 | $v$ | $v$ | $d(x,u) = d(u,v)$, $d(y,u) = d(u,v)$ |

We discuss here only the first two cases; the remaining cases are similar and are left to the reader.

In the first case, by Corollary 10.8, we have $d(x,u) \leq d(x,v) = d(u,v)$ and $d(y,u) \leq d(y,v) = d(u,v)$. This allows us to write

$$\begin{aligned}
&\max\{d(x,u) + d(y,v), d(x,v) + d(y,v)\} \\
&= \max\{d(x,u) + d(u,v), d(u,v) + d(y,v)\} \\
&= \max\{d(x,u), d(y,v)\} + d(u,v) \\
&\geq \max\{d(x,u), d(u,y)\} + d(u,v) \\
&\quad \text{(because } u \text{ is closer to } y \text{ than } v\text{)} \\
&\geq d(x,y) + d(u,v) \\
&\quad \text{(since } d \text{ is an ultrametric)},
\end{aligned}$$

which concludes the argument for the first case.

For the second case, by the same theorem mentioned above, we have $d(x,u) \leq d(x,v) = d(u,v)$ and $d(y,v) \leq d(y,u) = d(u,v)$. This implies that

$$d(x,u) + d(y,v) \leq d(x,v) + d(y,v) = 2d(u,v).$$

Thus, it remains to show only that $d(x,y) \leq d(u,v)$. Observe that we have $d(x,u) \leq d(u,v) = d(u,y)$. Therefore, in the triangle $x, y, u$, we have $d(x,y) = d(u,y) = d(u,v)$, which concludes the argument in the second case. □

Theorem 10.15 implies that for every set $S$, $\mathcal{U}_S \subseteq \mathcal{T}_S \subseteq \mathcal{M}_S$.

As shown in [24], Buneman's inequality is also a sufficient condition for a graph to be a tree in the following sense.

**Theorem 10.16.** *A graph $\mathcal{G} = (V, E)$ is a tree if and only if it is connected, contains no triangles, and its graph distance satisfies Buneman's inequality.*

*Proof.* By our previous discussions, the conditions are clearly necessary. We show here that they are sufficient.

Let $\mathbf{p}$ be a cycle of minimal length $\ell$. Since $\mathcal{G}$ contains no triangles, it follows that $\ell \geq 4$. Therefore, $\ell$ can be written as $\ell = 4q + r$, where $q \geq 1$ and $0 \leq r \leq 3$. Since $\mathbf{p}$ is a minimal circuit, the distance between its end points is given by the least number of edges of the circuit that separate the points. Therefore, we can select vertices $x, u, y, v$ (in this order) on the cycle such that the distances $d(x,u), d(u,y), d(y,v), d(v,x)$ are all either $q$ or $q+1$ and

$$d(x,u) + d(u,y) + d(y,v) + d(v,x) = 4q + r.$$

Then, $2q \leq d(x,y) \leq 2q+2$ and $2q \leq d(u,v) \leq 2q+2$, so $4q \leq d(x,y) + d(u,v) \leq 4q+4$, which prevents $d$ from satisfying the inequality $d(x,y) + d(u,v) \leq \max\{d(x,u) + d(y,v), d(x,v) + d(y,u)\}$. This condition shows that $\mathcal{G}$ is acyclic, so it is a tree. □

In data mining applied in biology, particularly in reconstruction of phylogenies, it is important to determine the conditions that allow the construction of a weighted tree $(\mathcal{T}, w)$ starting from a metric space $(S, d)$ such that the tree metric induced by $(\mathcal{T}, w)$ coincides with $d$ when restricted to the set $S$.

*Example 10.17.* Let $S = \{a, b, c\}$ be a three-element set and let $d$ be a distance defined on $S$. Suppose that $(a, b)$ are the closest points in $S$, that is, $d(a,b) \leq d(a,c)$ and $d(a,b) \leq d(b,c)$.

We shall seek to determine a weighted tree $(\mathcal{T}, w)$ such that the restriction of the metric induced by the tree to the set $S$ coincides with $d$. To this end, consider the weighted tree shown in Figure 10.3. The distances between vertices can be expressed as

$$\begin{aligned} d(a,b) &= m + n, \\ d(a,c) &= m + p + q, \\ d(b,c) &= n + p + q. \end{aligned}$$

It is easy to see that

$$p + q = \frac{d(a,c) + d(b,c) - d(a,b)}{2}.$$

A substitution in the last two equalities yields

$$\begin{aligned} m &= \frac{d(a,c) - d(b,c) + d(a,b)}{2} \geq 0, \\ n &= \frac{d(b,c) - d(a,c) + d(a,b)}{2} \geq 0, \end{aligned}$$

which determines the weights of the edges that end in $a$ and $b$, respectively. For the remaining two edges, one can choose $p$ and $q$ as two arbitrary positive numbers whose sum equals $\frac{d(a,c)+d(b,c)-d(a,b)}{2}$.

**Theorem 10.18.** *Starting from a tree metric $d$ on a nonempty set $S$, there exists a weighted tree $(\mathcal{T}, w)$ whose set of vertices contains $S$ and such that the metric induced by this weighted tree on $S$ coincides with $d$.*

*Proof.* The argument is by induction on $n = |S|$. The basis step, $n = 3$, is immediate.

Suppose that the statement holds for sets with fewer than $n$ elements, and let $S$ be a set with $|S| = n$. Define a function $f : S^3 \longrightarrow \mathbb{R}$ as $f(x,y,z) =$

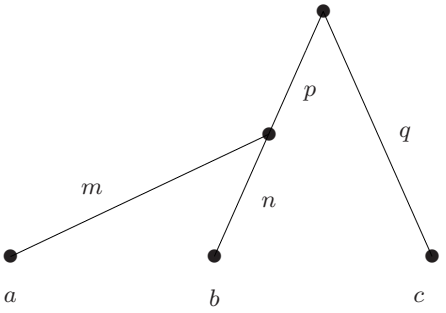



**Fig. 10.3.** Weighted tree.

$d(x,z) + d(y,z) - d(x,y)$. Let $(p,q,r) \in S^3$ be a triple such that $f(p,q,r)$ is maximum. If $x \in S - \{p,q\}$, we have $f(x,q,r) \leq f(p,q,r)$ and $f(p,x,r) \leq f(p,q,r)$. These inequalities are easily seen to be equivalent to

$$d(x,r) + d(p,q) \leq d(x,q) + d(p,r),$$
$$d(x,r) + d(p,q) \leq d(x,p) + d(q,r),$$

respectively. Using Buneman's inequality, we obtain

$$d(x,q) + d(p,r) = d(x,p) + d(q,r). \tag{10.2}$$

Similarly, for any other $y \in S - \{p,q\}$, we have

$$d(y,q) + d(p,r) = d(y,p) + d(q,r),$$

so

$$d(x,q) + d(y,p) = d(x,p) + d(y,q).$$

Consider now a new object $t$, $t \notin S$. The distances from $t$ to the objects of $S$ are defined by

$$d(t,p) = \frac{d(p,q) + d(p,r) - d(q,r)}{2},$$

and

$$d(t,x) = d(x,p) - d(t,p), \tag{10.3}$$

where $x \neq p$.

For $x \neq p$, we can write

$$\begin{aligned} d(t,x) &= d(x,p) - d(t,p) \\ &= d(x,p) - \frac{d(p,q) + d(p,r) - d(q,r)}{2} \\ &\quad \text{(by the definition of } d(t,p)) \\ &= d(x,p) - \frac{d(p,q) + d(p,x) - d(q,x)}{2} \\ &\quad \text{(by Equality (10.2))} \\ &= \frac{d(p,x) - d(p,q) + d(q,x)}{2} \geq 0. \end{aligned}$$

Choosing $x = q$ in Equality (10.3), we have

$$\begin{aligned} d(t,q) &= d(q,p) - d(t,p) \\ &= d(p,q) - \frac{d(p,q) + d(p,r) - d(q,r)}{2} \\ &= \frac{d(p,q) - d(p,r) + d(q,r)}{2} \geq 0, \end{aligned}$$

which shows that the distances of the form $d(t,\cdot)$ are all nonnegative.

Also, we can write for $x \in S - \{p,q\}$

$$\begin{aligned} d(q,t) + d(t,x) &= \frac{d(p,q) - d(p,r) + d(q,r)}{2} + \frac{d(p,x) - d(p,q) + d(q,x)}{2} \\ &= \frac{d(p,q) - d(p,x) + d(q,x)}{2} + \frac{d(p,x) - d(p,q) + d(q,x)}{2} \\ &\quad \text{(by Equality (10.2))} \\ &= d(q,x). \end{aligned}$$

It is not difficult to verify that the expansion of $d$ to $S \cup \{t\}$ using the values defined above satisfies Buneman's inequality.

Consider the metric space $((S - \{p,q\}) \cup \{t\}, d)$ defined over a set with $n-1$ elements. By inductive hypothesis, there exists a weighted tree $(\mathcal{T}, w)$ such that the metric induced on $(S - \{p,q\}) \cup \{t\}$ coincides with $d$. Adding two edges $(t,p)$ and $(t,q)$ having the weights $d(t,p)$ and $d(t,q)$, we obtain a tree that generates the distance $d$ on the set $S$. □

A class of weighted trees that is useful in clustering algorithms and in phylogenetics is introduced next.

**Definition 10.19.** *An* equidistant tree *is a triple* $(\mathcal{T}, w, v_0)$*, where* $(\mathcal{T}, v_0)$ *is a rooted tree and* $w$ *is a weighting function defined on the set of edges of* $\mathcal{T}$ *such that* $d_w(v_0, v)$ *is the same for every leaf of the rooted tree* $(\mathcal{T}, v_0)$*.*

In an equidistant tree $(\mathcal{T}, w; v_0)$, for every vertex $u$ there is a number $k$ such that $d_w(u,t) = k$ for every leaf that is joined to $v_0$ by a path that passes through $u$. In other words, the equidistant property is inherited by subtrees.

*Example 10.20.* The tree shown in Figure 10.4 is an equidistant tree. The distance $d_w$ from the root to each of its four leaves is equal to 8.

**Theorem 10.21.** *A function* $d : S \times S \longrightarrow \mathbb{R}$ *is an ultrametric if and only if there exists an equidistant tree* $(\mathcal{T}, w; v_0)$ *having* $S$ *as its set of leaves and* $d$ *is the restriction of the tree distance* $d_w$ *to* $S$*.*

*Proof.* To prove that the condition is necessary, let $(\mathcal{T}, w; v_0)$ be an equidistant tree and let $x, y, z \in L$ be three leaves of the tree. Suppose that $u$ is the

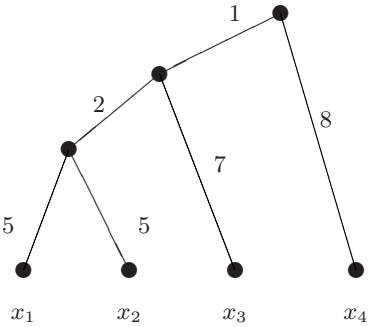



**Fig. 10.4.** An equidistant tree.

common ancestor of $x$ and $y$ located on the shortest path that joins $x$ to $y$. Then, $d_w(x, y) = 2d_w(u, x) = 2d_w(u, y)$.

Let $v$ be the common ancestor of $y$ and $z$ located on the shortest path that joins $y$ to $z$. Since both $u$ and $v$ are ancestors of $y$, they are located on the path that joins $v_0$ to $y$. Two cases may occur:

`Case 1` occurs when $d_w(v_0, v) \leq d_w(v_0, u)$ (Figure 10.5(a)).

`Case 2` occurs when $d_w(v_0, v) > d_w(v_0, u)$ (Figure 10.5(b)).

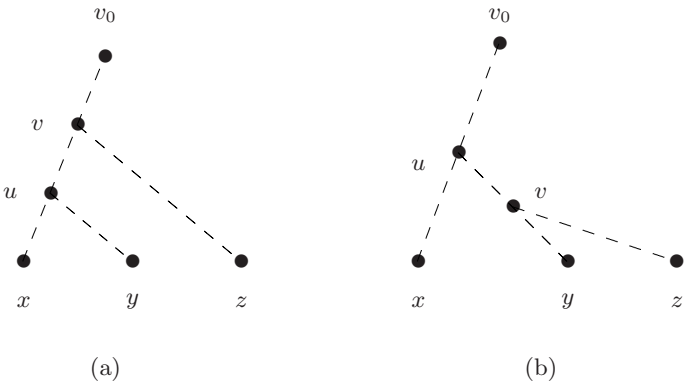



**Fig. 10.5.** Two equidistant trees.

In the first case, we have $d(u, x) = d(u, y)$ and $d(v, z) = d(v, u) + d(u, x) = d(v, u) + d(u, y)$ because $(\mathcal{T}, w, v_0)$ is equidistant. Therefore,

$$\begin{aligned} d_w(x, y) &= 2d_w(x, u), \\ d_w(y, z) &= 2d_w(u, v) + 2d_w(u, y), \\ d_w(x, z) &= 2d_w(u, v) + 2d_w(u, x). \end{aligned}$$

Since $d_w(u, y) = d_w(u, x)$, the ultrametric inequality

$$d_w(x, y) \leq \max\{d_w(x, z), d_w(z, y)\}$$

follows immediately. The second case is similar and is left to the reader.

Conversely, let $d : S \times S \longrightarrow \mathbb{R}$ be an ultrametric, where $S = \{s_1, \ldots, s_n\}$. We prove by induction of $n = |S|$ that an equidistant tree can be constructed that satisfies the requirements of the theorem.

For $n = 2$, the simple tree shown in Figure 10.6(a), where $w(x_0, s_1) = w(x_0, s_2) = \frac{d(s_1,s_2)}{2}$ satisfies the requirements of the theorem. For $n = 3$, suppose that $d(s_1, s_2) \leq d(s_1, s_3) = d(s_2, s_3)$. The tree shown in Figure 10.6(b) is the desired tree for the ultrametric because $d(s_1, s_3) - d(s_1, s_2) \geq 0$.

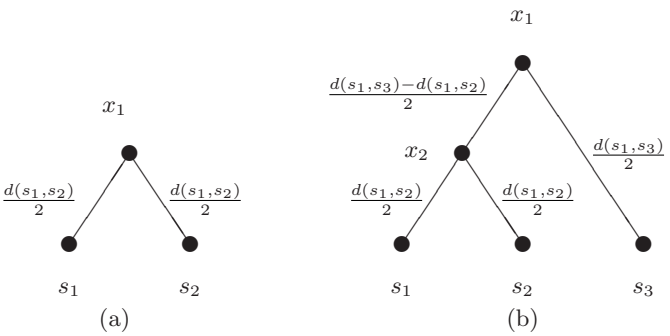



**Fig. 10.6.** Small equidistant weighted trees.

Suppose now that $n \geq 4$. Let $s_i, s_j$ be a pair of elements of $S$ such that the distance $d(s_i, s_j)$ is minimal. By the ultrametric property, we have $d(s_k, s_i) = d(s_k, s_j) \geq d(s_i, s_j)$ for every $k \in \{1, \ldots, n\} - \{i, j\}$.

Define $S' = S - \{s_i, s_j\} \cup \{s\}$, and let $d' : S' \times S' \longrightarrow \mathbb{R}$ be the mapping given by

$$d'(s_k, s_l) = d(s_k, s_l) \text{ if } s_k, s_l \in S$$

and $d'(s_k, s) = d(s_k, s_i) = d(s_k, s_j)$. It is easy to see that $d'$ is an ultrametric on the smaller set $S'$, so, by inductive hypothesis, there exists an equidistant weighted tree $(\mathcal{T}', w'; v_0)$ that induces $d'$ on the set of its leaves $S'$.

Let $z$ be the direct ancestor of $s$ in the tree $\mathcal{T}'$ and let $s_m$ be a neighbor of $s$. The weighted rooted tree $(\mathcal{T}, w; v_0)$ is obtained from $(\mathcal{T}', w'; v_0)$ by transforming $s$ into an interior node that has the leaves $s_i$ and $s_j$ as immediate descendants, as shown in Figure 10.7. To make the new tree $\mathcal{T}$ be equidistant, we keep all weights of the edges of $\mathcal{T}'$ in the new tree $\mathcal{T}$ except the weight of the edge $(z, s)$, which is defined now as

$$w(z, s) = \frac{d(s_m, s_i) - d(s_i, s_j)}{2}.$$

We also define the weight of the edges $(s, s_i)$ and $(s, s_j)$ as

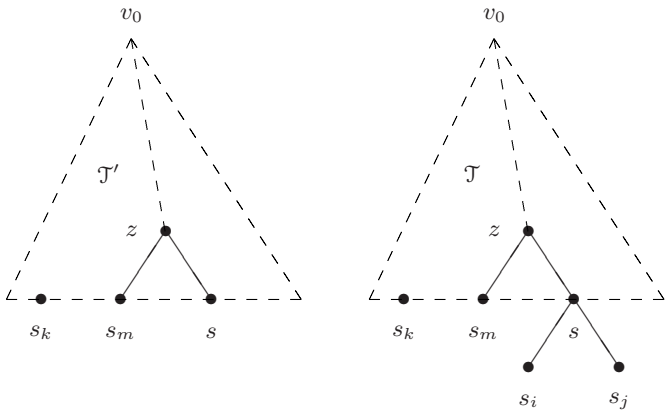



**Fig. 10.7.** Constructing $(\mathcal{T}, w; v_0)$ starting from $(\mathcal{T}', w'; v_0)$.

$$w(s, s_i) = w(s, s_j) = \frac{d(s_i, s_j)}{2}.$$

These definitions imply that $\mathcal{T}$ is an equidistant tree because

$$\begin{aligned} d'(s_m, z) &= \frac{d(s_m, s_i)}{2} \\ &\quad \text{(because of the definition of } d\text{)} \\ &= d'(z, s) \\ &\quad \text{(since } \mathcal{T}' \text{ is equidistant),} \end{aligned}$$

$$d(z, s_i) = w(z, s) + w(s, s_i) = \frac{d(s_m, s_i) - d(s_i, s_j)}{2} + \frac{d(s_i, s_j}{2} = \frac{d(s_m, s_i)}{2}.$$

and $d(z, z_m) = d'(z, s_m)$. □

## 10.4 Ultrametric Spaces

In Section 10.2, we saw that ultrametrics represent a strengthening of the notion of a metric, where the triangular inequality is replaced by a stronger requirement.

**Theorem 10.22.** *Let $B(x, r)$ be a closed sphere in the ultrametric space $(S, d)$. If $z \in B(x, d)$, then $B(x, r) = B(z, r)$. In other words, in an ultrametric space, a closed sphere has all its points as centers.*

*Proof.* Suppose that $z \in B(x, r)$, so $d(x, z) \leq r$. We saw that two of the numbers $d(x, z), d(z, y), d(x, y)$ are equal and the third is less than the equal numbers for any $y \in S$.

Let $y \in B(z,r)$. Since $d(y,x) \leq \max\{d(y,z), d(z,x)\} \leq r$, we have $y \in B(x,r)$. Conversely, if $y \in B(x,r)$, we have $d(y,z) \leq \max d(d(y,x), d(x,z) \leq r$, hence $y \in B(z,r)$. □

**Corollary 10.23.** *If two closed spheres $B(x,r)$ and $B(y,r')$ of an ultrametric space have a point in common, then one of the closed spheres is included in the other.*

*Proof.* The statement follows directly from Theorem 10.22. □

Theorem 10.22 implies that the entire space $S$ equals the closed sphere $B(x, diam_{S,d})$ for any point $x \in S$.

### Ultrametrics, Partitions, and Equivalences

We present now the link between ultrametrics defined on a finite set $S$ and chains of equivalence relations on $S$ (or chains of partitions on $S$). The next statement gives a method of constructing ultrametrics starting from chains of equivalence relations.

**Theorem 10.24.** *Let $S$ be a finite set and let $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ be a function whose range is* $\mathrm{Ran}(f) = \{r_1, \ldots, r_m\}$*, where $r_1 = 0$ such that $d(x,y) = 0$ if and only if $x = y$. Define the relations $\eta_{r_i} = \{(x,y) \in S \times S \mid d(x,y) \leq r_i\}$ for $1 \leq i \leq m$.*

*The function $d$ is an ultrametric on $S$ if and only if the sequence of relations $\eta_{r_1}, \ldots, \eta_{r_m}$ is an increasing sequence of equivalences on $S$ such that $\eta_{r_1} = \iota_S$ and $\eta_{r_m} = \theta_S$.*

*Proof.* Suppose that $d$ is an ultrametric on $S$. We have $(x,x) \in \eta_{r_i}$ because $d(x,x) = 0$, so all relations $\eta_{r_i}$ are reflexive. Also, it is clear that the symmetry of $d$ implies $(x,y) \in \eta_{r_i}$ if and only if $(y,x) \in \eta_{r_i}$, so these relations are symmetric.

The ultrametric inequality is essential for proving the transitivity of the relations $\eta_{r_i}$. If $(x,y), (y,z) \in \eta_{r_i}$, then $d(x,y) \leq r_i$ and $d(y,z) \leq r_i$, which implies $d(x,z) \leq \max\{d(x,y), d(y,z)\} \leq r_i$. Thus, $(x,z) \in \eta_{r_i}$, which shows that every relation $\eta_{r_i}$ is transitive and therefore an equivalence.

It is straightforward to see that $\eta_{r_1} \leq \eta_{r_2} \leq \cdots \leq \eta_{r_m}$; that is, this sequence of relations is indeed a chain of equivalences.

Conversely, suppose that $\eta_{r_1}, \ldots, \eta_{r_m}$ is an increasing sequence of equivalences on $S$ such that $\eta_{r_1} = \iota_S$ and $\eta_{r_m} = \theta_S$, where $\eta_{r_i} = \{(x,y) \in S \times S \mid d(x,y) \leq r_i\}$ for $1 \leq i \leq m$ and $r_1 = 0$.

Note that $d(x,y) = 0$ is equivalent to $(x,y) \in \eta_{r_1} = \iota_S$, that is, to $x = y$.

We claim that

$$d(x,y) = \min\{r \mid (x,y) \in \eta_r\}. \tag{10.4}$$

Indeed, since $\eta_{r_m} = \theta_S$, it is clear that there is an equivalence $\eta_{r_i}$ such that $(x,y) \in \eta_{r_i}$. If $(x,y) \in \eta_{r_i}$, the definition of $\eta_{r_i}$ implies $d(x,y) \leq r_i$, so

$d(x, y) \leq \min\{r \mid (x, y) \in \eta_r\}$. This inequality can be easily seen to become an equality since $(x, y) \in \eta_{d(x,y)}$. This implies immediately that $d$ is symmetric.

To prove that $d$ satisfies the ultrametric inequality, let $x, y, z$ be three members of the set $S$. Let $p = \max\{d(x, z), d(z, y)\}$. Since $(x, z) \in \eta_{d(x,z)} \subseteq \eta_p$ and $(z, y) \in \eta_{d(z,y)} \subseteq \eta_p$, it follows that $(x, y) \in \eta_p$, due to the transitivity of the equivalence $\eta_p$. Thus, $d(x, y) \leq p = \max\{d(x, z), d(z, y)\}$, which proves the triangular inequality for $d$. □

Of course, Theorem 10.24 can be formulated in terms of partitions.

**Theorem 10.25.** *Let $S$ be a finite set and let $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ be a function whose range is* $\mathrm{Ran}(f) = \{r_1, \ldots, r_m\}$*, where $r_1 = 0$ such that $d(x, y) = 0$ if and only if $x = y$. For $u \in S$ and $r \in \mathbb{R}_{\geq 0}$, define the set $D_{u,r} = \{x \in S \mid d(u, x) \leq r\}$.*

*Define the collection of sets $\pi_{r_i} = \{D(u, r_i) \mid u \in S\}$ for $1 \leq i \leq m$.*

*The function $d$ is an ultrametric on $S$ if and only if the sequence of collections $\pi_{r_1}, \ldots, \pi_{r_m}$ is an increasing sequence of partitions on $S$ such that $\pi_{r_1} = \alpha_S$ and $\pi_{r_m} = \omega_S$.*

*Proof.* The argument is entirely similar to the proof of Theorem 10.24 and is omitted. □

### Hierarchies and Ultrametrics

**Definition 10.26.** *Let $S$ be a set. A* hierarchy on the set $S$ *is a collection of sets $\mathcal{H} \subseteq \mathcal{P}(S)$ that satisfies the following conditions:*

(i) *the members of $\mathcal{H}$ are nonempty sets;*
(ii) *$S \in \mathcal{H}$;*
(iii) *for every $x \in S$, we have $\{x\} \in \mathcal{H}$;*
(iv) *if $H, H' \in \mathcal{H}$ and $H \cap H' \neq \emptyset$, then we have either $H \subseteq H'$ or $H' \subseteq H$.*

A standard technique for constructing a hierarchy on a set $S$ starts with a rooted tree $(\mathcal{T}, v_0)$ whose nodes are labeled by subsets of the set $S$. Let $V$ be the set of vertices of the tree $\mathcal{T}$. The function $\mu : V \longrightarrow \mathcal{P}(S)$, which gives the label $\mu(v)$ of each node $v \in V$, is defined as follows:

(i) The tree $\mathcal{T}$ has $|S|$ leaves, and each leaf $v$ is labeled by a distinct singleton $\mu(v) = \{x\}$ for $x \in S$.
(ii) If an interior vertex $v$ of the tree has the descendants $v_1, v_2, \ldots, v_n$, then $\mu(v) = \bigcup_{i=1}^{n} \mu(v_i)$.

The set of labels $\mathcal{H}_{\mathcal{T}}$ of the rooted tree $(\mathcal{T}, v_0)$ forms a hierarchy on $S$. Indeed, note that each singleton $\{x\}$ is a label of a leaf. An easy argument by induction on the height of the tree shows that every vertex is labeled by the set of labels of the leaves that descend from that vertex. Therefore, the root $v_0$ of the tree is labeled by $S$.

Suppose that $H, H'$ are labels of the nodes $u, v$ of $\mathcal{T}$, respectively. If $H \cap H' \neq \emptyset$, then the vertices $u, v$ have a common descendant. In a tree, this can

take place only if $u$ is a descendant of $v$ or $v$ is a descendant of $u$; that is, only if $H \subseteq H'$, or $H' \subseteq H$, respectively. This gives the desired conclusion.

*Example 10.27.* Let $S = \{s, t, u, v, w, x, y\}$ and let $\mathcal{T}$ be a tree whose vertices are labeled as shown in Figure 10.8. It is easy to verify that the family of subsets of $S$ that label the nodes of $\mathcal{T}$,

$$\begin{aligned}\mathcal{H} = \{&\{s\}, \{t\}, \{u\}, \{v\}, \{w\}, \{x\}, \{y\},\\ &\{s, t, u\}, \{w, x\}, \{s, t, u, v\}, \{w, x, y\}, \{s, t, u, v, w, x, y\}\}\end{aligned}$$

is a hierarchy on the set $S$.

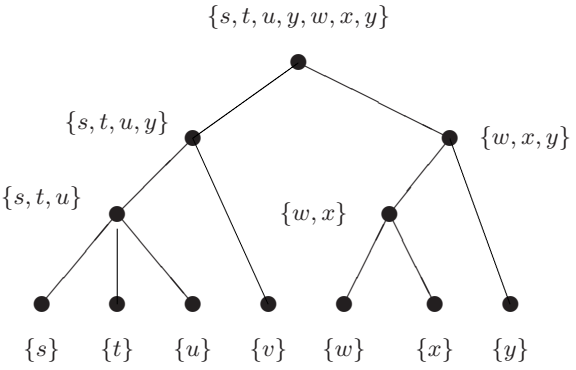



**Fig. 10.8.** Tree labeled by subsets of $S$.

Chains of partitions defined on a set generate hierarchies, as we show next.

**Theorem 10.28.** *Let $S$ be a set and let $C = (\pi_1, \pi_2, \ldots, \pi_n)$ be an increasing chain of partitions $(PART(S), \leq)$ such that $\pi_1 = \alpha_S$ and $\pi_n = \omega_S$. Then, the collection $\mathcal{H}_C = \bigcup_{i=1}^{n} \pi_i$ that consists of the blocks of all partitions in the chain is a hierarchy on $S$.*

*Proof.* The blocks of any of the partitions are nonempty sets, so $\mathcal{H}_C$ satisfies the first condition of Definition 10.26.

We have $S \in \mathcal{H}_C$ because $S$ is the unique block of $\pi_n = \omega_S$. Also, since all singletons $\{x\}$ are blocks of $\alpha_S = \pi_1$, it follows that $\mathcal{H}_C$ satisfies the second and the third conditions of Definition 10.26. Finally, let $H$ and $H'$ be two sets of $\mathcal{H}_C$ such that $H \cap H' \neq \emptyset$. Because of this condition, it is clear that these two sets cannot be blocks of the same partition. Thus, there exist two partitions $\pi_i$ and $\pi_j$ in the chain such that $H \in \pi_i$ and $H' \in \pi_j$. Suppose that $i < j$. Since every block of $\pi_j$ is a union of blocks of $\pi_i$, $H'$ is a union of blocks of $\pi_i$ and $H \cap H' \neq \emptyset$ means that $H$ is one of these blocks. Thus, $H \subseteq H'$. If

$j > i$, we obtain the reverse inclusion. This allows us to conclude that $\mathcal{H}_C$ is indeed a hierarchy. □

Theorem 10.28 can be stated in terms of chains of equivalences; we give the following alternative formulation for convenience.

**Theorem 10.29.** *Let $S$ be a finite set and let $(\rho_1, \ldots, \rho_n)$ be a chain of equivalence relations on $S$ such that $\rho_1 = \iota_S$ and $\rho_n = \theta_S$. Then, the collection of blocks of the equivalence relations $\rho_r$ (that is, the set $\bigcup_{1 \leq r \leq n} S/\rho_r$) is a hierarchy on $S$.*

*Proof.* The proof is a mere restatement of the proof of Theorem 10.28. □

Define the relation "$\prec$" on a hierarchy $\mathcal{H}$ on $S$ by $H \prec K$ if $H, K \in \mathcal{H}$, $H \subset K$, and there is no set $L \in \mathcal{H}$ such that $H \subset L \subset K$.

**Lemma 10.30.** *Let $\mathcal{H}$ be a hierarchy on a finite set $S$ and let $L \in \mathcal{H}$. The collection $\mathcal{P}_L = \{H \in \mathcal{H} \mid H \prec L\}$ is a partition of the set $L$.*

*Proof.* We claim that $L = \bigcup \mathcal{P}_L$. Indeed, it is clear that $\bigcup \mathcal{P}_L \subseteq L$.

Conversely, suppose that $z \in L$ but $z \notin \bigcup \mathcal{P}_L$. Since $\{z\} \in \mathcal{H}$ and there is no $K \in \mathcal{P}_L$ such that $z \in K$, it follows that $\{z\} \in \mathcal{P}_L$, which contradicts the assumption that $z \notin \bigcup \mathcal{P}_L$. This means that $L = \bigcup \mathcal{P}_L$.

Let $K_0, K_1 \in \mathcal{P}_L$ be two distinct sets. These sets are disjoint since otherwise we would have either $K_0 \subset K_1$ or $K_1 \subset K_0$, and this would contradict the definition of $\mathcal{P}_L$. □

**Theorem 10.31.** *Let $\mathcal{H}$ be a hierarchy on a set $S$. The graph of the relation $\prec$ on $\mathcal{H}$ is a tree whose root is $S$; its leaves are the singletons $\{x\}$ for every $x \in S$.*

*Proof.* Since $\prec$ is an antisymmetric relation on $\mathcal{H}$, it is clear that the graph $(\mathcal{H}, \prec)$ is acyclic. Moreover, for each set $K \in \mathcal{H}$, there is a unique path that joins $K$ to $S$, so the graph is indeed a rooted tree. □

**Definition 10.32.** *Let $\mathcal{H}$ be a hierarchy on a set $S$. A* grading function *for $\mathcal{H}$ is a function $h : \mathcal{H} \longrightarrow \mathbb{R}$ that satisfies the following conditions:*
(i) *$h(\{x\}) = 0$ for every $x \in S$, and*
(ii) *if $H, K \in \mathcal{H}$ and $H \subset K$, then $h(H) < h(K)$.*

*If $h$ is a grading function for a hierarchy $\mathcal{H}$, the pair $(\mathcal{H}, h)$ is a* graded hierarchy.

*Example 10.33.* For the hierarchy $\mathcal{H}$ defined in Example 10.27 on the set $S = \{s, t, u, v, w, x, y\}$, the function $h : \mathcal{H} \longrightarrow \mathbb{R}$ given by

$$h(\{s\}) = h(\{t\}) = h(\{u\}) = h(\{v\}) = h(\{w\}) = h(\{x\}) = h(\{y\}) = 0,$$
$$h(\{s, t, u\}) = 3, h(\{w, x\}) = 4, h(\{s, t, u, v\}) = 5, h(\{w, x, y\}) = 6,$$
$$h(\{s, t, u, v, w, x, y\}) = 7,$$

is a grading function and the pair $(\mathcal{H}, h)$ is a graded hierarchy on $S$.

Theorem 10.28 can be extended to graded hierarchies.

**Theorem 10.34.** *Let $S$ be a finite set and let $C = (\pi_1, \pi_2, \ldots, \pi_n)$ be an increasing chain of partitions $(PART(S), \leq)$ such that $\pi_1 = \alpha_S$ and $\pi_n = \omega_S$.*

*Consider a function $f : \{1, \ldots, n\} \longrightarrow \mathbb{R}_{\geq 0}$ such that $f(1) = 0$. The function $h : \mathcal{H}_C \longrightarrow \mathbb{R}_{\geq 0}$ given by*

$$h(K) = f\left(\min\{j \mid K \in \pi_j\}\right)$$

*is a grading function for the hierarchy $\mathcal{H}_C$.*

*Proof.* Since $\{x\} \in \pi_1 = \alpha_S$, it follows that $h(\{x\}) = 0$, so $h$ satisfies the first condition of Definition 10.32.

Suppose that $H, K \in \mathcal{H}_C$ and $H \subset K$. If $\ell = \min\{j \mid H \in \pi_j\}$ it is impossible for $K$ to be a block of a partition that precedes $\pi_\ell$. Therefore, $\ell < \min\{j \mid K \in \pi_j\}$, so $h(H) < h(K)$, and $(\mathcal{H}_C, h)$ is indeed a graded hierarchy. □

A graded hierarchy defines an ultrametric, as shown next.

**Theorem 10.35.** *Let $(\mathcal{H}, h)$ be a graded hierarchy on a finite set $S$. Define the function $d : S^2 \longrightarrow \mathbb{R}$ as*

$$d(x, y) = \min\{h(U) \mid U \in \mathcal{H} \text{ and } \{x, y\} \subseteq U\}$$

*for $x, y \in S$. The mapping $d$ is an ultrametric on $S$.*

*Proof.* Observe that for every $x, y \in S$ there exists a set $H \in \mathcal{H}$ such that $\{x, y\} \subseteq H$ because $S \in \mathcal{H}$.

It is immediate that $d(x, x) = 0$. Conversely, suppose that $d(x, y) = 0$. Then, there exists $H \in \mathcal{H}$ such that $\{x, y\} \subseteq H$ and $h(H) = 0$. If $x \neq y$, then $\{x\} \subset H$, hence $0 = h(\{x\}) < h(H)$, which contradicts the fact that $h(H) = 0$. Thus, $x = y$.

The symmetry of $d$ is immediate.

To prove the ultrametric inequality, let $x, y, z \in S$, and suppose that $d(x, y) = p$, $d(x, z) = q$, and $d(z, y) = r$. There exist $H, K, L \in \mathcal{H}$ such that $\{x, y\} \subseteq H$, $h(H) = p$, $\{x, z\} \subseteq K$, $h(K) = q$, and $\{z, y\} \subseteq L$, $h(L) = r$. Since $K \cap L \neq \emptyset$ (because both sets contain $z$), we have either $K \subseteq L$ or $L \subseteq K$, so $K \cup L$ equals either $K$ or $L$ and, in either case, $K \cup L \in \mathcal{H}$. Since $\{x, y\} \subseteq K \cup L$, it follows that

$$d(x, y) \leq h(K \cup L) = \max\{h(K), H(L)\} = \max\{d(x, z), d(z, y)\},$$

which is the ultrametric inequality. □

We refer to the ultrametric $d$ whose existence is shown in Theorem 10.35 as the *ultrametric generated by the graded hierarchy* $(\mathcal{H}, h)$.

*Example 10.36.* The values of the ultrametric generated by the graded hierarchy $(\mathcal{H}, h)$ on the set $S$ introduced in Example 10.33 are given in the following table:

| $d$ | $s$ | $t$ | $u$ | $v$ | $w$ | $x$ | $y$ |
|---|---|---|---|---|---|---|---|
| $s$ | 0 | 3 | 3 | 5 | 7 | 7 | 7 |
| $t$ | 3 | 0 | 3 | 5 | 7 | 7 | 7 |
| $u$ | 3 | 3 | 0 | 5 | 7 | 7 | 7 |
| $v$ | 5 | 5 | 5 | 0 | 7 | 7 | 7 |
| $w$ | 7 | 7 | 7 | 7 | 0 | 4 | 6 |
| $x$ | 7 | 7 | 7 | 7 | 4 | 0 | 6 |
| $y$ | 7 | 7 | 7 | 7 | 6 | 6 | 0 |

The hierarchy introduced in Theorem 10.29 that is associated with an ultrametric space can be naturally equipped with a grading function, as shown next.

**Theorem 10.37.** *Let $(S, d)$ be a finite ultrametric space. There exists a graded hierarchy $(\mathcal{H}, h)$ on $S$ such that $d$ is the ultrametric associated to $(\mathcal{H}, h)$.*

*Proof.* Let $\mathcal{H}$ be the collection of equivalence classes of the equivalences $\eta_r = \{(x, y) \in S^2 \mid d(x, y) \leq r\}$ defined by the ultrametric $d$ on the finite set $S$, where the index $r$ takes its values in the range $R_d$ of the ultrametric $d$. Define $h(E) = \min\{r \in R_d \mid E \in S/\eta_r\}$ for every equivalence class $E$.

It is clear that $h(\{x\}) = 0$ because $\{x\}$ is an $\eta_0$-equivalence class for every $x \in S$.

Let $[x]_t$ be the equivalence class of $x$ relative to the equivalence $\eta_t$.

Suppose that $E$ and $E'$ belong to the hierarchy and $E \subset E'$. We have $E = [x]_r$ and $E' = [x]_s$ for some $x \in X$. Since $E$ is strictly included in $E'$, there exists $z \in E' - E$ such that $d(x, z) \leq s$ and $d(x, z) > r$. This implies $r < s$. Therefore,

$$h(E) = \min\{r \in R_d \mid E \in S/\eta_r\} \leq \min\{s \in R_d \mid E' \in S/\eta_s\} = h(E'),$$

which proves that $(\mathcal{H}, h)$ is a graded hierarchy.

The ultrametric $e$ generated by the graded hierarchy $(\mathcal{H}, h)$ is given by

$$\begin{aligned} e(x, y) &= \min\{h(B) \mid B \in \mathcal{H} \text{ and } \{x, y\} \subseteq B\} \\ &= \min\{r \mid (x, y) \in \eta_r\} \\ &= \min\{r \mid d(x, y) \leq r\} \\ &= d(x, y), \end{aligned}$$

for $x, y \in S$; in other words, we have $e = d$. □

*Example 10.38.* Starting from the ultrametric on the set $S = \{s, t, u, v, w, x, y\}$ defined by the table given in Example 10.36, we obtain the following quotient sets:

| Values of $r$ | $S/\eta_r$ |
|---|---|
| $[0, 3)$ | $\{s\}, \{t\}, \{u\}, \{v\}, \{w\}, \{x\}, \{y\}$ |
| $[3, 4)$ | $\{s, t, u\}, \{v\}, \{w\}, \{x\}, \{y\}$ |
| $[4, 5)$ | $\{s, t, u\}, \{v\}, \{w, x\}, \{y\}$ |
| $[5, 6)$ | $\{s, t, u, v\}, \{w, x\}, \{y\}$ |
| $[6, 7)$ | $\{s, t, u, v\}, \{w, x, y\}$ |
| $[7, \infty)$ | $\{s, t, u, v, w, x, y\}$ |

We shall draw the tree of a graded hierarchy $(\mathcal{H}, h)$ using a special representation known as a *dendrogram*. In a dendrogram, an interior vertex $K$ of the tree is represented by a horizontal line drawn at the height $h(K)$. For example, the dendrogram of the graded hierarchy of Example 10.33 is shown in Figure 10.9.

As we saw in Theorem 10.35, the value $d(x, y)$ of the ultrametric $d$ generated by a hierarchy $\mathcal{H}$ is the smallest height of a set of a hierarchy that contains both $x$ and $y$. This allows us to "read" the value of the ultrametric generated by $\mathcal{H}$ directly from the dendrogram of the hierarchy.

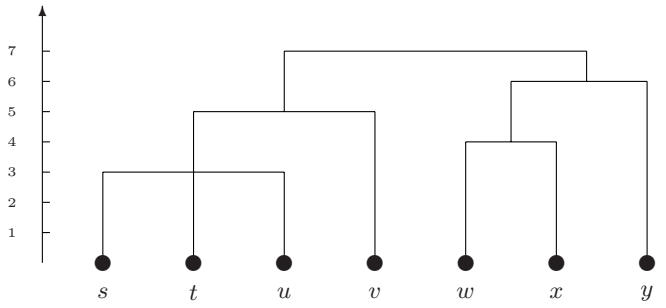



**Fig. 10.9.** Dendrogram of graded hierarchy of Example 10.33.

*Example 10.39.* For the graded hierarchy of Example 10.33, the ultrametric extracted from Figure 10.9 is clearly the same as the one that was obtained in Example 10.36.

### The Poset of Ultrametrics

Let $S$ be a set. Recall that we denoted the set of dissimilarities by $\mathcal{D}_S$. Define a partial order $\leq$ on $\mathcal{D}_S$ by $d \leq d'$ if $d(x, y) \leq d'(x, y)$ for every $x, y \in S$. It is easy to verify that $(\mathcal{D}_S, \leq)$ is a poset.

The set $\mathcal{U}_S$ of ultrametrics on $S$ is a subset of $\mathcal{D}_S$.

**Theorem 10.40.** *Let $d$ be a dissimilarity on a set $S$ and let $U_d$ be the set of ultrametrics*

$$U_d = \{e \in \mathcal{U}_S \mid e \leq d\}.$$

*The set $U_d$ has a largest element in the poset $(\mathcal{D}_S, \leq)$.*

*Proof.* The set $U_d$ is nonempty because the zero dissimilarity $d_0$ given by $d_0(x, y) = 0$ for every $x, y \in S$ is an ultrametric and $d_0 \leq d$.

Since the set $\{e(x, y) \mid e \in U_d\}$ has $d(x, y)$ as an upper bound, it is possible to define the mapping $e_1 : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ as

$$e_1(x, y) = \sup\{e(x, y) \mid e \in U_d\}$$

for $x, y \in S$. It is clear that $e \leq e_1$ for every ultrametric $e$. We claim that $e_1$ is an ultrametric on $S$.

We prove only that $e_1$ satisfies the ultrametric inequality. Suppose that there exist $x, y, z \in S$ such that $e_1$ violates the ultrametric inequality; that is,

$$\max\{e_1(x, z), e_1(z, y)\} < e_1(x, y).$$

This is equivalent to

$$\begin{aligned}&\sup\{e(x, y) \mid e \in U_d\}\\ &> \max\{\sup\{e(x, z) \mid e \in U_d\}, \sup\{e(z, y) \mid e \in U_d\}\}.\end{aligned}$$

Thus, there exists $\hat{e} \in U_d$ such that

$$\begin{aligned}\hat{e}(x, y) &> \sup\{e(x, z) \mid e \in U_d\},\\ \hat{e}(x, y) &> \sup\{e(z, y) \mid e \in U_d\}.\end{aligned}$$

In particular, $\hat{e}(x, y) > \hat{e}(x, z)$ and $\hat{e}(x, y) > \hat{e}(z, y)$, which contradicts the fact that $\hat{e}$ is an ultrametric. □

The ultrametric defined by Theorem 10.40 is known as the *maximal subdominant ultrametric for the dissimilarity $d$*.

The situation is not symmetric with respect to the infimum of a set of ultrametrics because, in general, the infimum of a set of ultrametrics is not necessarily an ultrametric.

For example, consider a three-element set $S = \{x, y, z\}$, four distinct nonnegative numbers $a, b, c, d$ such that $a > b > c > d$ and the ultrametrics $d$ and $d'$ defined by the triangles shown in Figures 10.10(a) and (b), respectively. The dissimilarity $d_0$ defined by $d_0(u, v) = \min\{d(u, v), d'(u, v)\}$ for $u, v \in S$ is given by

$$d_0(x, y) = b, d_0(y, z) = d, \text{ and } d_0(x, z) = c,$$

and $d_0$ is clearly not an ultrametric because the triangle $xyz$ is not isosceles.

In what follows, we give an algorithm for computing the maximal subdominant ultrametric for a dissimilarity defined on a finite set $S$.

We will define inductively an increasing sequence of partitions $\pi_1 \prec \pi_2 \prec \cdots$ and a sequence of dissimilarities $d_1, d_2, \ldots$ on the sets of blocks of $\pi_1, \pi_2, \ldots$, respectively.

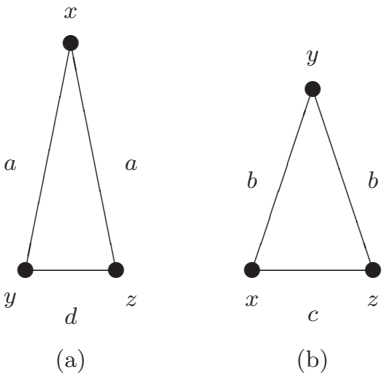



**Fig. 10.10.** Two ultrametrics on the set $\{x, y, z\}$.

For the initial phase, $\pi_1 = \alpha_S$ and $d_1(\{x\}, \{y\}) = d(x, y)$ for $x, y \in S$.

Suppose that $d_i$ is defined on $\pi_i$. If $B, C \in \pi_i$ is a pair of blocks such that $d_i(B, C)$ has the smallest value, define the partition $\pi_{i+1}$ by

$$\pi_{i+1} = (\pi_i - \{B, C\}) \cup \{B \cup C\}.$$

In other words, to obtain $\pi_{i+1}$, we replace two of the closest blocks $B$ and $C$, of $\pi_i$ (in terms of $d_i$) with new block $B \cup C$. Clearly, $\pi_i \prec \pi_{i+1}$ in $PART(S)$ for $i \geq 1$. Note that the collection of blocks of the partitions $\pi_i$ forms a hierarchy $\mathcal{H}_d$ on the set $S$. The dissimilarity $d_{i+1}$ is given by

$$d_{i+1}(U, V) = \min\{d(x, y) \mid x \in U, y \in V\} \tag{10.5}$$

for $U, V \in \pi_{i+1}$.

We introduce a grading function $h_d$ on the hierarchy defined by this chain of partitions starting from the dissimilarity $d$. The definition is done for the blocks of the partitions $\pi_i$ by induction on $i$.

For $i = 1$ the blocks of the partition $\pi_1$ are singletons; in this case we define $h_d(\{x\}) = 0$ for $x \in S$.

Suppose that $h_d$ is defined on the blocks of $\pi_i$, and let $D$ be the block of $\pi_{i+1}$ that is generated by fusing the blocks $B$ and $C$ of $\pi_i$. All other blocks of $\pi_{i+1}$ coincide with the blocks of $\pi_i$. The value of the function $h_d$ for the new block $D$ is given by

$$h_d(D) = \min\{d(x, y) \mid x \in B, y \in C\}.$$

It is clear that $h_d$ satisfies the first condition of Definition 10.32.

For a set $U$ of $\mathcal{H}_d$, define

$$p_U = \min\{i \mid U \in \pi_i\} \text{ and } q_U = \max\{i \mid U \in \pi_i\}.$$

To verify the second condition of Definition 10.32, let $H, K \in \mathcal{H}_d$ such that $H \subset K$. It is clear that $q_H \leq p_K$. The construction of the sequence of partitions implies that there are $H_0, H_1 \in \pi_{p_H - 1}$ and $K_0, K_1 \in \pi_{p_K - 1}$ such that $H = H_0 \cup H_1$ and $K = K_0 \cup K_1$. Therefore,

$$h_d(H) = \min\{d(x, y) \mid x \in H_0, y \in H_1\},$$
$$h_d(K) = \min\{d(x, y) \mid x \in K_0, y \in K_1\}.$$

Since $H_0$ and $H_1$ were fused (to produce the partition $\pi_{p_H}$) before $K_0$ and $K_1$ were (to produce the partition $\pi_{p_K}$), it follows that $h_d(H) < h_d(K)$.

By Theorem 10.35, the graded hierarchy $(\mathcal{H}_d, h_d)$ defines an ultrametric; we denote this ultrametric by $e$ and will prove that $e$ is the maximal subdominant ultrametric for $d$. Recall that $e$ is given by

$$e(x, y) = \min\{h_d(W) \mid \{x, y\} \subseteq W\}$$

and that $h_d(W)$ is the least value of $d(u, v)$ such that $u \in U, v \in V$ if $W \in \pi_{p_W}$ is obtained by fusing the blocks $U$ and $V$ of $\pi_{p_W - 1}$. The definition of $e(x, y)$ implies that we have neither $\{x, y\} \subseteq U$ nor $\{x, y\} \subseteq V$. Thus, we have either $x \in U$ and $y \in V$ or $x \in V$ and $y \in U$. Thus, $e(x, y) \leq d(x, y)$.

We now prove that:

$$e(x, y) = \min\{amp_d(\mathbf{s}) \mid \mathbf{s} \in S(x, y)\}$$

for $x, y \in S$.

Let $D$ be the minimal set in $\mathcal{H}_d$ that includes $\{x, y\}$. Then, $D = B \cup C$, where $B$ and $C$ are two disjoint sets of $\mathcal{H}_d$ such that $x \in B$ and $y \in C$. If $\mathbf{s}$ is a sequence included in $D$, then there are two consecutive components of $\mathbf{s}$, $s_k$ and $s_{k+1}$, such that $s_k \in B$ and $s_{k+1} \in C$. This implies

$$\begin{aligned} e(x, y) &= \min\{d(u, v) \mid u \in B, v \in C\} \\ &\leq d(s_k, s_{k+1}) \\ &\leq amp_d(\mathbf{s}). \end{aligned}$$

If $\mathbf{s}$ is not included in $D$, let $s_q$ and $s_{q+1}$ be two consecutive components of $\mathbf{s}$ such that $s_q \in D$ and $s_{q+1} \notin D$. Let $E$ be the smallest set of $\mathcal{H}_d$ that includes $\{s_q, s_{q+1}\}$. We have $D \subseteq E$ (because $s_k \in D \cap E$) and therefore $h_d(D) \leq h_d(E)$. If $E$ is obtained as the union of two disjoint sets $E'$ and $E''$ of $\mathcal{H}_d$ such that $s_k \in E'$ and $s_{k+1} \in E''$, we have $D \subseteq E'$. Consequently,

$$h_d(E) = \min\{d(u, v) \mid u \in E', v \in E''\} \leq d(s_k, s_{k+1}),$$

which implies

$$e(x, y) = h_d(D) \leq h_d(E) \leq d(s_k, s_{k+1}) \leq amp_d(\mathbf{s}).$$

Therefore, we conclude that $e(x, y) \leq amp_d(\mathbf{s})$ for every $\mathbf{s} \in S(x, y)$.

We now show that there is a sequence $\mathbf{w} \in S(x, y)$ such that $e(x, y) \geq amp_d(\mathbf{w})$, which implies the equality $e(x, y) = amp_d(\mathbf{w})$. To this end, we prove that for every $D \in \pi_k \subseteq \mathcal{H}_d$ there exists $\mathbf{w} \in S(x, y)$ such that $amp_d(\mathbf{w}) \leq h_d(D)$. The argument is by induction on $k$.

For $k = 1$, the statement obviously holds. Suppose that it holds for $1, \ldots, k-1$, and let $D \in \pi_k$. The set $D$ belongs to $\pi_{k-1}$ or $D$ is obtained by fusing the blocks $B, C$ of $\pi_{k-1}$. In the first case, the statement holds by inductive hypothesis. The second case has several subcases:

(i) If $\{x, y\} \subseteq B$, then by the inductive hypothesis, there exists a sequence $\mathbf{u} \in S(x, y)$ such that $amp_d(\mathbf{u}) \leq h_d(B) \leq h_d(D) = e(x, y)$.

(ii) The case $\{x, y\} \subseteq C$ is similar to the first case.

(iii) If $x \in B$ and $y \in C$, there exist $u, v \in D$ such that $d(u, v) = h_d(D)$. By the inductive hypothesis, there is a sequence $\mathbf{u} \in S(x, u)$ such that $amp_d(\mathbf{u}) \leq h_d(B)$ and there is a sequence $\mathbf{v} \in S(v, y)$ such that $amp_d(\mathbf{v}) \leq h_d(C)$. This allows us to consider the sequence $\mathbf{w}$ obtained by concatenating the sequences $\mathbf{u}, (u, v), \mathbf{v}$; clearly, $\mathbf{w} \in S(x, y)$ and $amp_d(\mathbf{w}) = \max\{amp_d(\mathbf{u}), d(u, v), amp_d(\mathbf{v})\} \leq h_d(D)$.

To complete the argument, we need to show that if $e'$ is another ultrametric such that $e(x, y) \leq e'(x, y) \leq d(x, y)$, then $e(x, y) = e'(x, y)$ for every $x, y \in S$. By the previous argument, there exists a sequence $\mathbf{s} = (s_0, \ldots, s_n) \in S(x, y)$ such that $amp_d(\mathbf{s}) = e(x, y)$. Since $e'(x, y) \leq d(x, y)$ for every $x, y \in S$, it follows that $e'(x, y) \leq amp_d(\mathbf{s}) = e(x, y)$. Thus, $e(x, y) = e'(x, y)$ for every $x, y \in S$, which means that $e = e'$. This concludes our argument.

## 10.5 Metrics on $\mathbb{R}^n$

Data sets often consist of $n$-dimensional vectors having real number components. Dissimilarities between these vectors can be evaluated by using one of the metrics that we present in this section. We begin with two technical results.

**Lemma 10.41.** *Let $p, q \in \mathbb{R} - \{0, 1\}$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then we have $p > 1$ if and only if $q > 1$. Furthermore, one of the numbers $p, q$ belongs to the interval $(0, 1)$ if and only if the other number is negative.*

*Proof.* We leave to the reader the simple proof of this statement. □

**Lemma 10.42.** *Let $p, q \in \mathbb{R} - \{0, 1\}$ be two numbers such that $\frac{1}{p} + \frac{1}{q} = 1$ and $p > 1$. Then, for every $a, b \in \mathbb{R}_{\geq 0}$, we have*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q},$$

*where the equality holds if and only if $a = b^{-\frac{1}{1-p}}$.*

*Proof.* By Lemma 10.41, we have $q > 1$. Consider the function $f(x) = \frac{x^p}{p} + \frac{1}{q} - x$ for $x \geq 0$. We have $f'(x) = x^{p-1} - 1$, so the minimum is achieved when $x = 1$ and $f(1) = 0$. Thus,

$$f\left(ab^{-\frac{1}{p-1}}\right) \geq f(1) = 0,$$

which amounts to

$$\frac{a^p b^{-\frac{p}{p-1}}}{p} + \frac{1}{q} - ab^{-\frac{1}{p-1}} \geq 0.$$

By multiplying both sides of this inequality by $b^{\frac{p}{p-1}}$, we obtain the desired inequality. □

Observe that if $\frac{1}{p} + \frac{1}{q} = 1$ and $p < 1$, then $q < 0$. In this case, we have the reverse inequality

$$ab \geq \frac{a^p}{p} + \frac{b^q}{q}. \tag{10.6}$$

which can be shown by observing that the function $f$ has a maximum in $x = 1$. The same inequality holds when $q < 1$ and therefore $p < 0$.

**Theorem 10.43 (The Hölder Inequality).** *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be $2n$ nonnegative numbers, and let $p$ and $q$ be two numbers such that $\frac{1}{p} + \frac{1}{q} = 1$ and $p > 1$. We have*

$$\sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} b_i^q\right)^{\frac{1}{q}}.$$

*Proof.* Define the numbers

$$x_i = \frac{a_i}{(\sum_{i=1}^{n} a_i^p)^{\frac{1}{p}}} \text{ and } y_i = \frac{b_i}{(\sum_{i=1}^{n} b_i^q)^{\frac{1}{q}}}$$

for $1 \leq i \leq n$. Lemma 10.42 applied to $x_i, y_i$ yields

$$\frac{a_i b_i}{(\sum_{i=1}^{n} a_i^p)^{\frac{1}{p}} (\sum_{i=1}^{n} b_i^q)^{\frac{1}{q}}} \leq \frac{1}{p} \frac{a_i^p}{\sum_{i=1}^{n} a_i^p} + \frac{1}{q} \frac{b_i^p}{\sum_{i=1}^{n} b_i^p}.$$

Adding these inequalities, we obtain

$$\sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} b_i^q\right)^{\frac{1}{q}}$$

because $\frac{1}{p} + \frac{1}{q} = 1$. □

The nonnegativity of the numbers $a_1, \ldots, a_n, b_1, \ldots, b_n$ can be relaxed by using absolute values. Indeed, we can easily prove the following variant of Theorem 10.43.

**Theorem 10.44.** *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be $2n$ numbers and let $p$ and $q$ be two numbers such that $\frac{1}{p} + \frac{1}{q} = 1$ and $p > 1$. We have*

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leq \left( \sum_{i=1}^{n} |a_i|^p \right)^{\frac{1}{p}} \cdot \left( \sum_{i=1}^{n} |b_i|^q \right)^{\frac{1}{q}}.$$

*Proof.* By Theorem 10.43, we have

$$\sum_{i=1}^{n} |a_i||b_i| \leq \left( \sum_{i=1}^{n} |a_i|^p \right)^{\frac{1}{p}} \cdot \left( \sum_{i=1}^{n} |b_i|^q \right)^{\frac{1}{q}}.$$

The needed equality follows from the fact that

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leq \sum_{i=1}^{n} |a_i||b_i|.$$

□

**Corollary 10.45 (The Cauchy Inequality).** *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be $2n$ numbers. We have*

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leq \sqrt{\sum_{i=1}^{n} |a_i|^2} \cdot \sqrt{\sum_{i=1}^{n} |b_i|^2}.$$

*Proof.* The inequality follows immediately from Theorem 10.44 by taking $p = q = 2$. □

**Theorem 10.46 (Minkowski's Inequality).** *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be $2n$ nonnegative numbers. If $p \geq 1$, we have*

$$\left( \sum_{i=1}^{n} (a_i + b_i)^p \right)^{\frac{1}{p}} \leq \left( \sum_{i=1}^{n} a_i^p \right)^{\frac{1}{p}} + \left( \sum_{i=1}^{n} b_i^p \right)^{\frac{1}{p}}.$$

*If $p < 1$, the inequality sign is reversed.*

*Proof.* For $p = 1$, the inequality is immediate. Therefore, we can assume that $p > 1$. Note that

$$\sum_{i=1}^{n} (a_i + b_i)^p = \sum_{i=1}^{n} a_i (a_i + b_i)^{p-1} + \sum_{i=1}^{n} b_i (a_i + b_i)^{p-1}.$$

By Hölder's inequality for $p, q$ such that $p > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, we have

$$\sum_{i=1}^{n} a_i(a_i+b_i)^{p-1} \leq \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n}(a_i+b_i)^{(p-1)q}\right)^{\frac{1}{q}}$$
$$= \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n}(a_i+b_i)^{p}\right)^{\frac{1}{q}}.$$

Similarly, we can write

$$\sum_{i=1}^{n} b_i(a_i+b_i)^{p-1} \leq \left(\sum_{i=1}^{n} b_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n}(a_i+b_i)^{p}\right)^{\frac{1}{q}}.$$

Adding the last two inequalities yields

$$\sum_{i=1}^{n}(a_i+b_i)^p \leq \left(\left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} b_i^p\right)^{\frac{1}{p}}\right) \left(\sum_{i=1}^{n}(a_i+b_i)^{p}\right)^{\frac{1}{q}},$$

which is equivalent to the desired inequality

$$\left(\sum_{i=1}^{n}(a_i+b_i)^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} b_i^p\right)^{\frac{1}{p}}.$$

□

**Corollary 10.47.** *For $p \geq 1$, the function $\nu_p : \mathbb{R}^n \longrightarrow \mathbb{R}_{\geq 0}$ defined by*

$$\nu_p(x_1, \ldots, x_n) = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}},$$

*where $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, is a norm on the linear space $(\mathbb{R}^n, +, \cdot)$.*

*Proof.* We must prove that $\nu_p$ satisfies the conditions of Definition 2.47. The argument for the first two parts of this definition are immediate and are left to the reader.

Let $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Minkowski's inequality applied to the nonnegative numbers $a_i = |x_i|$ and $b_i = |y_i|$ amounts to

$$\left(\sum_{i=1}^{n}(|x_i|+|y_i|)^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{\frac{1}{p}}.$$

Since $|x_i + y_i| \leq |x_i| + |y_i|$ for every $i$, we have

$$\left(\sum_{i=1}^{n}(|x_i+y_i|)^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{\frac{1}{p}},$$

that is, $\nu_p(\mathbf{x}+\mathbf{y}) \leq \nu_p(\mathbf{x}) + \nu_p(\mathbf{y})$. Thus, $\nu_p$ is a norm on $\mathbb{R}^n$. □

*Example 10.48.* Consider the mappings $\nu_1, \nu_\infty : \mathbb{R}^n \longrightarrow \mathbb{R}$ given by

$$\nu_1(\mathbf{x}) = |x_1| + |x_2| + \cdots + |x_n|,$$
$$\nu_\infty(\mathbf{x}) = \max\{|x_1|, |x_2|, \ldots, |x_n|\},$$

for every $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Both $\nu_1$ and $\nu_\infty$ are norms on $\mathbb{R}^n$. However, neither $\nu_1$ nor $\nu_\infty$ are generated by an inner product on $\mathbb{R}^n$ (see Exercise 19).

We will frequently use the alternative notation $\| \mathbf{x} \|_p$ for $\nu_p(\mathbf{x})$.

A special metric on $\mathbb{R}^n$ is the function $\nu_\infty : \mathbb{R}^n \longrightarrow \mathbb{R}_{\geq 0}$ given by

$$\nu_\infty(\mathbf{x}) = \max\{|x_i| \mid 1 \leq i \leq n\} \tag{10.7}$$

for $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

$\nu_\infty$ clearly satisfies the first two conditions of Definition 2.47. To prove that it satisfies the third condition, we start from the inequality

$$|x_i + y_i| \leq |x_i| + |y_i| \leq \nu_\infty(\mathbf{x}) + \nu_\infty(\mathbf{y})$$

for every $i$, $1 \leq i \leq n$. This in turn implies

$$\nu_\infty(\mathbf{x} + \mathbf{y}) = \max\{|x_i + y_i| \mid 1 \leq i \leq n\} \leq \nu_\infty(\mathbf{x}) + \nu_\infty(\mathbf{y}),$$

which gives the desired inequality.

This norm can be regarded as a limit case of the norms $\nu_p$. Indeed, let $\mathbf{x} \in \mathbb{R}^n$ and let $M = \max\{|x_i| \mid 1 \leq i \leq n\} = |x_{\ell_1}| = \cdots = |x_{\ell_k}|$ for some $\ell_1, \ldots, \ell_k$, where $1 \leq \ell_1, \ldots, \ell_k \leq n$. Here $x_{\ell_1}, \ldots, x_{\ell_k}$ are the components of $\mathbf{x}$ that have the maximal absolute value and $k \geq 1$. We can write

$$\lim_{p \to \infty} \nu_p(\mathbf{x}) = \lim_{p \to \infty} M \left( \sum_{i=1}^{n} \left( \frac{|x_i|}{M} \right)^p \right)^{\frac{1}{p}} = \lim_{p \to \infty} M(k)^{\frac{1}{p}} = M,$$

which justifies the notation $\nu_\infty$.

The following statement holds for every linear space and therefore for the linear space $(\mathbb{R}^n, +, \cdot)$.

**Theorem 10.49.** *Each norm $\nu : L \longrightarrow \mathbb{R}_{\geq 0}$ on a metric space $(L, +, \cdot)$ generates a metric on the set $L$ defined by $d_\nu(\boldsymbol{x}, \boldsymbol{y}) = \| \boldsymbol{x} - \boldsymbol{y} \|$ for $\boldsymbol{x}, \boldsymbol{y} \in L$.*

*Proof.* Note that, by the first property of norms from Definition 2.47, if $d_\nu(\mathbf{x}, \mathbf{y}) = \| \mathbf{x} - \mathbf{y} \| = 0$, it follows that $\mathbf{x} - \mathbf{y} = \mathbf{0}$; that is, $\mathbf{x} = \mathbf{y}$.

The symmetry of $d_\nu$ is obvious and so we need to verify only the triangular axiom. Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in L$. Applying the third property of Definition 2.47, we have

$$\nu(\mathbf{x} - \mathbf{z}) = \nu(\mathbf{x} - \mathbf{y} + \mathbf{y} - \mathbf{z}) \leq \nu(\mathbf{x} - \mathbf{y}) + \nu(\mathbf{y} - \mathbf{z})$$

or, equivalently, $d_\nu(\mathbf{x}, \mathbf{z}) \leq d_\nu(\mathbf{x}, \mathbf{y}) + d_\nu(\mathbf{y}, \mathbf{z})$, for every $\mathbf{x}, \mathbf{y}, \mathbf{z} \in L$, which concludes the argument. □

We refer to $d_\nu$ as the *metric induced by the norm $\nu$ on the linear space* $(L, +, \cdot)$.

For $p \geq 1$, then $d_p$ denotes the metric $d_{\nu_p}$ induced by the norm $\nu_p$ on the linear space $(\mathbb{R}^n, +, \cdot)$ known as the *Minkowski metric* on $\mathbb{R}^n$.

If $p = 2$, we have the *Euclidean metric* on $\mathbb{R}^n$ given by

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

For $p = 1$, we have

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|.$$

These metrics can be seen in Figure 10.11 for the special case of $\mathbb{R}^2$. If $\mathbf{x} = (x_0, x_1)$ and $\mathbf{y} = (y_0, y_1)$, then $d_2(\mathbf{x}, \mathbf{y})$ is the length of the hypotenuse of the right triangle and $d_1(\mathbf{x}, \mathbf{y})$ is the sum of the lengths of the two legs of the triangle.

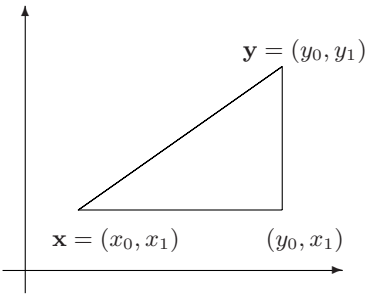



**Fig. 10.11.** The distances $d_1(\mathbf{x}, \mathbf{y})$ and $d_2(\mathbf{x}, \mathbf{y})$.

Theorem 10.51 to follow allows us to compare the norms $\nu_p$ (and the metrics of the form $d_p$) that were introduced on $\mathbb{R}^n$. We begin with a preliminary result.

**Lemma 10.50.** *Let $a_1, \ldots, a_n$ be $n$ positive numbers. If $p$ and $q$ are two positive numbers such that $p \leq q$, then*

$$(a_1^p + \cdots + a_n^p)^{\frac{1}{p}} \geq (a_1^q + \cdots + a_n^q)^{\frac{1}{q}}.$$

*Proof.* Let $f : \mathbb{R}^{>0} \longrightarrow \mathbb{R}$ be the function defined by

$$f(r) = (a_1^r + \cdots + a_n^r)^{\frac{1}{r}}.$$

Since

$$\ln f(r) = \frac{\ln (a_1^r + \cdots + a_n^r)}{r},$$

it follows that

$$\frac{f'(r)}{f(r)} = -\frac{1}{r^2}(a_1^r + \cdots + a_n^r) + \frac{1}{r} \cdot \frac{a_1^r \ln a_1 + \cdots + a_n^r \ln a_r}{a_1^r + \cdots + a_n^r}.$$

To prove that $f'(r) < 0$, it suffices to show that

$$\frac{a_1^r \ln a_1 + \cdots + a_n^r \ln a_r}{a_1^r + \cdots + a_n^r} \leq \frac{\ln (a_1^r + \cdots + a_n^r)}{r}.$$

This last inequality is easily seen to be equivalent to

$$\sum_{i=1}^{n} \frac{a_i^r}{a_1^r + \cdots + a_n^r} \ln \frac{a_i^r}{a_1^r + \cdots + a_n^r} \leq 0,$$

which holds because

$$\frac{a_i^r}{a_1^r + \cdots + a_n^r} \leq 1$$

for $1 \leq i \leq n$. □

**Theorem 10.51.** *Let $p$ and $q$ be two positive numbers such that $p \leq q$. For every $\boldsymbol{u} \in \mathbb{R}^n$, we have $\| \boldsymbol{u} \|_p \geq \| \boldsymbol{u} \|_q$.*

*Proof.* This statement follows immediately from Lemma 10.50. □

**Corollary 10.52.** *Let $p, q$ be two positive numbers such that $p \leq q$. For every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have $d_p(\boldsymbol{x}, \boldsymbol{y}) \geq d_q(\boldsymbol{x}, \boldsymbol{y})$.*

*Proof.* This statement follows immediately from Theorem 10.51. □

**Theorem 10.53.** *Let $p \geq 1$. For every $\boldsymbol{x} \in \mathbb{R}^n$ we have*

$$\| \boldsymbol{x} \|_\infty \leq \| \boldsymbol{x} \|_p \leq n \| \boldsymbol{x} \|_\infty .$$

*Proof.* The first inequality is an immediate consequence of Theorem 10.51. The second inequality follows by observing that

$$\| \mathbf{x} \|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \leq n \max_{1 \leq i \leq n} |x_i| = n \| \mathbf{x} \|_\infty .$$

□

**Corollary 10.54.** *Let $p$ and $q$ be two numbers such that $p, q \geq 1$. There exist two constants $c, d \in \mathbb{R}_{>0}$ such that*

$$c \parallel \boldsymbol{x} \parallel_q \leq \parallel \boldsymbol{x} \parallel_p \leq d \parallel \boldsymbol{x} \parallel_q$$

*for $\boldsymbol{x} \in \mathbb{R}^n$.*

*Proof.* Since $\parallel \mathbf{x} \parallel_\infty \leq \parallel \mathbf{x} \parallel_p$ and $\parallel x \parallel_q \leq n \parallel \mathbf{x} \parallel_\infty$, it follows that $\parallel x \parallel_q \leq n \parallel \mathbf{x} \parallel_p$. Exchanging the roles of $p$ and $q$, we have $\parallel x \parallel_p \leq n \parallel \mathbf{x} \parallel_q$, so

$$\frac{1}{n} \parallel \mathbf{x} \parallel_q \leq \parallel \mathbf{x} \parallel_p \leq n \parallel \mathbf{x} \parallel_q$$

for every $\mathbf{x} \in \mathbb{R}^n$. □

**Corollary 10.55.** *For every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $p \geq 1$, we have $d_\infty(\boldsymbol{x}, \boldsymbol{y}) \leq d_p(\boldsymbol{x}, \boldsymbol{y}) \leq n d_\infty(\boldsymbol{x}, \boldsymbol{y})$. Further, for $p, q > 1$, there exist $c, d \in \mathbb{R}_{>0}$ such that*

$$c d_q(\boldsymbol{x}, \boldsymbol{y}) \leq d_p(\boldsymbol{x}, \boldsymbol{y}) \leq c d_q(\boldsymbol{x}, \boldsymbol{y})$$

*for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.*

*Proof.* This follows from Theorem 10.53 and from Corollary 10.55. □

Corollary 10.52 implies that if $p \leq q$, then the closed sphere $B_{d_p}(\mathbf{x}, r)$ is included in the closed sphere $B_{d_q}(\mathbf{x}, r)$. For example, we have

$$B_{d_1}(\mathbf{0}, 1) \subseteq B_{d_2}(\mathbf{0}, 1) \subseteq B_{d_\infty}(\mathbf{0}, 1).$$

In Figures 10.12 (a) - (c) we represent the closed spheres $B_{d_1}(\mathbf{0}, 1)$, $B_{d_2}(\mathbf{0}, 1)$, and $B_{d_\infty}(\mathbf{0}, 1)$.

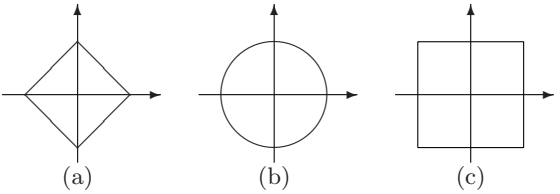



**Fig. 10.12.** Spheres $B_{d_p}(\mathbf{0}, 1)$ for $p = 1, 2, \infty$.

An useful consequence of Theorem 10.43 is the following statement:

**Theorem 10.56.** *Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_m$ be $2m$ nonnegative numbers such that $\sum_{i=1}^m x_i = \sum_{i=1}^m y_i = 1$ and let $p$ and $q$ be two positive numbers such that $\frac{1}{p} + \frac{1}{q} = 1$. We have*

$$\sum_{j=1}^m x_j^{\frac{1}{p}} y_j^{\frac{1}{q}} \leq 1.$$

*Proof.* The Hölder inequality applied to $x_1^{\frac{1}{p}}, \ldots, x_m^{\frac{1}{p}}$ and $y_1^{\frac{1}{q}}, \ldots, y_m^{\frac{1}{q}}$ yields the needed inequality

$$\sum_{j=1}^{m} x_j^{\frac{1}{p}} y_j^{\frac{1}{q}} \leq \sum_{j=1}^{m} x_j \sum_{j=1}^{m} y_j = 1$$

□

Theorem 10.56 allows the formulation of a generalization of the Hölder Inequality.

**Theorem 10.57.** *Let $\boldsymbol{A}$ be an $n \times m$ matrix, $\boldsymbol{A} = (a_{ij})$, having positive entries such that $\sum_{j=1}^{m} a_{ij} = 1$ for $1 \leq i \leq n$. If $\boldsymbol{p} = (p_1, \ldots, p_n)$ is an $n$-tuple of positive numbers such that $\sum_{i=1}^{n} p_i = 1$, then*

$$\sum_{j=1}^{m} \prod_{i=1}^{n} a_{ij}^{p_i} \leq 1.$$

*Proof.* The argument is by induction on $n \geq 2$. The basis case, $n = 2$ follows immediately from Theorem 10.56 by choosing $p = \frac{1}{p_1}$, $q = \frac{1}{p_2}$, $x_j = a_{1j}$, and $y_j = a_{2j}$ for $1 \leq j \leq m$.

Suppose that the statement holds for $n$, let $\mathbf{A}$ be an $(n+1) \times m$-matrix having positive entries such that $\sum_{j=1}^{m} a_{ij} = 1$ for $1 \leq i \leq n+1$, and let $\mathbf{p} = (p_1, \ldots, p_n, p_{n+1})$ be such that $p_1 + \cdots + p_n + p_{n+1} = 1$.

It is easy to see that

$$\sum_{j=1}^{m} \prod_{i=1}^{n+1} a_{ij}^{p_i} \leq \sum_{j=1}^{m} a_{1j}^{p_1} a_{n-1\,j}^{p_{n-1}} (a_{nj} + a_{n+1\,j})^{p_n + p_{n+1}}.$$

By applying the inductive hypothesis, we have

$$\sum_{j=1}^{m} \prod_{i=1}^{n+1} a_{ij}^{p_i} \leq 1.$$

□

A more general form of Theorem 10.57 is given next.

**Theorem 10.58.** *Let $\boldsymbol{A}$ be an $n \times m$ matrix, $\boldsymbol{A} = (a_{ij})$, having positive entries. If $\boldsymbol{p} = (p_1, \ldots, p_n)$ is an $n$-tuple of positive numbers such that $\sum_{i=1}^{n} p_i = 1$, then*

$$\sum_{j=1}^{m} \prod_{i=1}^{n} a_{ij}^{p_i} \leq \prod_{i=1}^{n} \left( \sum_{j=1}^{m} a_{ij} \right)^{p_i}.$$

*Proof.* Let $\mathbf{B} = (b_{ij})$ be the matrix defined by

$$b_{ij} = \frac{a_{ij}}{\sum_{j=1}^{m} a_{ij}}$$

for $1 \leq i \leq n$ and $1 \leq j \leq m$. Since $\sum_{j=1}^{m} b_{ij} = 1$, we can apply Theorem 10.57 to this matrix. Thus, we can write

$$\begin{aligned}\sum_{j=1}^{m}\prod_{i=1}^{n} b_{ij}^{p_i} &= \sum_{j=1}^{m}\prod_{i=1}^{n}\left(\frac{a_{ij}}{\sum_{j=1}^{m} a_{ij}}\right)^{p_i}\\ &= \sum_{j=1}^{m}\prod_{i=1}^{n}\frac{a_{ij}^{p_i}}{\left(\sum_{j=1}^{m} a_{ij}\right)^{p_i}}\\ &= \frac{\sum_{j=1}^{m}\prod_{i=1}^{n} a_{ij}^{p_i}}{\prod_{i=1}^{n}\left(\sum_{j=1}^{m} a_{ij}\right)^{p_i}} \leq 1.\end{aligned}$$

□

We now give a generalization of Minkowski's inequality (Theorem 10.46). First, we need a preliminary result.

**Lemma 10.59.** *If $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are positive numbers and $r < 0$, then*

$$\sum_{i=1}^{n} a_i^r b_i^{1-r} \geq \left(\sum_{i=1}^{n} a_i\right)^r \cdot \left(\sum_{i=1}^{n} b_i\right)^{1-r}.$$

*Proof.* Let $c_1, \ldots, c_n, d_1, \ldots, d_n$ be $2n$ positive numbers such that $\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} d_i = 1$. Inequality (10.6) applied to the numbers $a = c_i^{\frac{1}{p}}$ and $b = d_i{}^{\frac{1}{q}}$ yields:

$$c_i^{\frac{1}{p}} d_i^{\frac{1}{q}} \geq \frac{c_i}{p} + \frac{d_i}{q}.$$

Summing these inequalities produces the inequality

$$\sum_{i=1}^{n} c_i^{\frac{1}{p}} d_i^{\frac{1}{q}} \geq 1,$$

or

$$\sum_{i=1}^{n} c_i^r d_i^{1-r} \geq 1,$$

where $r = \frac{1}{p} < 0$. Choosing $c_i = \frac{a_i}{\sum_{i=1}^{n} a_i}$ and $d_i = \frac{b_i}{\sum_{i=1}^{n} b_i}$, we obtain the desired inequality. □

**Theorem 10.60.** *Let $\boldsymbol{A}$ be an $n \times m$ matrix, $\boldsymbol{A} = (a_{ij})$, having positive entries, and let $p$ and $q$ be two numbers such that $p > q$ and $p \neq 0$, $q \neq 0$. We have*

$$\left(\sum_{j=1}^{m}\left(\sum_{i=1}^{n} a_{ij}^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}} \geq \left(\sum_{i=1}^{n}\left(\sum_{j=1}^{m} a_{ij}^q\right)^{\frac{p}{q}}\right)^{\frac{1}{p}}.$$

*Proof.* Define

$$E = \left( \sum_{j=1}^{m} \left( \sum_{i=1}^{n} a_{ij}^{p} \right)^{\frac{q}{p}} \right)^{\frac{1}{q}},$$

$$F = \left( \sum_{i=1}^{n} \left( \sum_{j=1}^{m} a_{ij}^{q} \right)^{\frac{p}{q}} \right)^{\frac{1}{p}},$$

and $u_i = \sum_{j=1}^{m} a_{ij}^{q}$ for $1 \leq i \leq n$.

There are three distinct cases to consider related to the position of 0 relative to $p$ and $q$.

Suppose initially that $p > q > 0$. We have

$$\begin{aligned} F^p &= \textstyle\sum_{i=1}^{n} u_i^{\frac{p}{q}} & &= \textstyle\sum_{i=1}^{n} u_i u_i^{\frac{p}{q}-1} \\ &= \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^{q} u_i^{\frac{p}{q}-1} & &= \textstyle\sum_{j=1}^{m} \sum_{i=1}^{n} a_{ij}^{q} u_i^{\frac{p}{q}-1}. \end{aligned}$$

By applying the Hölder inequality, we have

$$\begin{aligned} \sum_{i=1}^{n} a_{ij}^{q} u_i^{\frac{p}{q}-1} &\leq \left( \sum_{i=1}^{n} (a_{ij}^{q})^{\frac{p}{q}} \right)^{\frac{q}{p}} \cdot \left( \sum_{i=1}^{n} (u_i^{\frac{p}{q}-1})^{\frac{p}{p-q}} \right)^{1-\frac{q}{p}} \\ &= \left( \sum_{i=1}^{n} a_{ij}^{p} \right)^{\frac{q}{p}} \cdot \left( \sum_{i=1}^{n} u_i^{\frac{p}{q}} \right)^{1-\frac{q}{p}}, \end{aligned} \tag{10.8}$$

which implies $F^p \leq E^q F^{p-q}$. This, in turn, gives $F^q \leq E^q$, which implies the generalized Minkowski inequality.

Suppose now that $0 > p > q$, so $0 < -p < -q$. Applying the generalized Minkowski inequality to the positive numbers $b_{ij} = \frac{1}{a_{ij}}$ gives the inequality

$$\left( \sum_{j=1}^{m} \left( \sum_{i=1}^{n} b_{ij}^{-q} \right)^{\frac{p}{q}} \right)^{-\frac{1}{p}} \geq \left( \sum_{i=1}^{n} \left( \sum_{j=1}^{m} b_{ij}^{-p} \right)^{\frac{q}{p}} \right)^{-\frac{1}{q}},$$

which is equivalent to

$$\left( \sum_{j=1}^{m} \left( \sum_{i=1}^{n} a_{ij}^{q} \right)^{\frac{p}{q}} \right)^{-\frac{1}{p}} \geq \left( \sum_{i=1}^{n} \left( \sum_{j=1}^{m} a_{ij}^{p} \right)^{\frac{q}{p}} \right)^{-\frac{1}{q}}.$$

A last transformation gives

$$\left(\sum_{j=1}^{m}\left(\sum_{i=1}^{n} a_{ij}^{q}\right)^{\frac{p}{q}}\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{n}\left(\sum_{j=1}^{m} a_{ij}^{p}\right)^{\frac{q}{p}}\right)^{\frac{1}{q}},$$

which is the inequality to be proven.

Finally, suppose that $p > 0 > q$. Since $\frac{q}{p} < 0$, Inequality (10.9) is replaced by the opposite inequality through the application of Lemma 10.59:

$$\sum_{i=1}^{n} a_{ij}^{q} u_i^{\frac{p}{q}-1} \geq \left(\sum_{i=1}^{n} a_{ij}^{p}\right)^{\frac{q}{p}} \cdot \left(\sum_{i=1}^{n} u_i^{\frac{p}{q}}\right)^{1-\frac{q}{p}}.$$

This leads to $F^p \geq E^q F^{p-q}$ or $F^q \geq E^q$. Since $q < 0$, this implies $F \leq E$. □

## 10.6 Metrics on Collections of Sets

Dissimilarities between subsets of finite sets have an intrinsic interest for data mining, where comparisons between sets of objects are frequent. Also, metrics defined on subsets can be transferred to metrics between binary sequences using the characteristic sequences of the subsets and thus become an instrument for studying binary data.

A very simple metric on $\mathcal{P}(S)$, the set of subsets of a finite set $S$ is given in the next theorem.

**Theorem 10.61.** *Let $S$ be a finite set. The mapping $\delta : (\mathcal{P}(S))^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by $\delta(X, Y) = |X \oplus Y|$ is a metric on $\mathcal{P}(S)$.*

*Proof.* The function $\delta$ is clearly symmetric and we have $\delta(X, Y) = 0$ if and only if $X = Y$. Therefore, we need to prove only the triangular inequality

$$|X \oplus Y| \leq |X \oplus Z| + |Z \oplus Y|$$

for every $X, Y, Z \in \mathcal{P}(S)$.

Since $X \oplus Y = (X \oplus Z) \oplus (Z \oplus Y)$, we have $|X \oplus Y| \leq |X \oplus Z| + |Z \oplus Y|$, which is precisely the triangular inequality for $\delta$. □

For $U, V \in \mathcal{P}(S)$, we have $0 \leq \delta(U, V) \leq |S|$, where $\delta(U, V) = |S|$ if and only if $V = S - U$.

**Lemma 10.62.** *Let $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ be a metric and let $u \in S$ be an element of the set $S$. Define the* Steinhaus transform of $d$ *as the mapping $d_u : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ given by*

$$d_u(x, y) = \begin{cases} 0 & \text{if } x = y = u \\ \frac{d(x,y)}{d(x,y)+d(x,u)+d(u,y)} & \text{otherwise.} \end{cases}$$

*Then, $d_u$ is a metric on $S$.*

*Proof.* It is easy to see that $d_u$ is symmetric and, further, that $d_u(x, y) = 0$ if and only if $x = y$.

To prove the triangular inequality, observe that $a \leq a'$ implies

$$\frac{a}{a+k} \leq \frac{a'}{a'+k}, \tag{10.9}$$

which holds for every positive numbers $a, a', k$. Then, we have

$$\begin{aligned}
d_u(x,y) &= \frac{d(x,y)}{d(x,y)+d(x,u)+d(u,y)} \\
&\leq \frac{d(x,z)+d(z,y)}{d(x,z)+d(z,y)+d(x,u)+d(u,y)} \\
&\quad \text{(by Inequality (10.9))} \\
&= \frac{d(x,z)}{d(x,z)+d(z,y)+d(x,u)+d(u,y)} \\
&\quad + \frac{d(z,y)}{d(x,z)+d(z,y)+d(x,u)+d(u,y)} \\
&\leq \frac{d(x,z)}{d(x,z)+d(z,y)+d(z,u)} + \frac{d(z,y)}{d(z,y)+d(z,u)+d(u,y)} \\
&= d_u(x,z) + d_u(z,y),
\end{aligned}$$

which is the desired triangular inequality. □

**Theorem 10.63.** *Let $S$ be a finite set. The function $d : \mathcal{P}(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by*

$$d(X,Y) = \frac{|X \oplus Y|}{|X \cup Y|}$$

*for $X, Y \in \mathcal{P}(S)$ is a metric on $\mathcal{P}(S)$.*

*Proof.* It is clear that $d$ is symmetric and that $d(X, Y) = 0$ if and only if $X = Y$. So, we need to prove only the triangular inequality. The mapping $\delta$ defined by $\delta(X, Y) = |X \oplus Y|$ is a metric on $\mathcal{P}(X)$, as we proved in Theorem 10.61. By Lemma 10.62, the mapping $\delta_\emptyset$ is also a metric on $\mathcal{P}(S)$. We have

$$\delta_\emptyset(X,Y) = \frac{|X \oplus Y|}{|X \oplus Y| + |X \oplus \emptyset| + |\emptyset \oplus Y|}.$$

Since $X \oplus \emptyset = X$, $\emptyset \oplus Y = Y$, we have

$$|X \oplus Y| + |X \oplus \emptyset| + |\emptyset \oplus Y| = |X \oplus Y| + |X| + |Y| = 2|X \cup Y|,$$

which means that $2\delta_\emptyset(X, Y) = d(X, Y)$ for every $X, Y \in \mathcal{P}(S)$. This implies that $d$ is indeed a metric. □

**Theorem 10.64.** *Let $S$ be a finite set. The function $d : \mathcal{P}(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by*

$$d(X, Y) = \frac{|X \oplus Y|}{|S| - |X \cap Y|}$$

*for $X, Y \in \mathcal{P}(S)$ is a metric on $\mathcal{P}(S)$.*

*Proof.* We only prove that $d$ satisfies the triangular axiom. The argument begins, as in Theorem 10.63, with the metric $\delta$. Again, by Lemma 10.62, the mapping $\delta_S$ is also a metric on $\mathcal{P}(S)$. We have

$$\begin{aligned}
\delta_S(X, Y) &= \frac{|X \oplus Y|}{|X \oplus Y| + |X \oplus S| + |S \oplus Y|} \\
&= \frac{|X \oplus Y|}{|X \oplus Y| + |S - X| + |S - Y|} \\
&= \frac{|X \oplus Y|}{|X \oplus Y| + |S - X| + |S - Y|} \\
&= \frac{|X \oplus Y|}{2(|S| - |X \cap Y|)}
\end{aligned}$$

because $|X \oplus Y| + |S - X| + |S - Y| = 2(|S| - |X \cap Y|)$, as the reader can easily verify. Therefore, $d(X, Y) = 2\delta_S(X, Y)$, which proves that $d$ is indeed a metric. □

A general mechanism for defining a metric on $\mathcal{P}(S)$, where $S$ is a finite set, $|S| = n$, can be introduced starting with two functions:

1. a *weight function* $w : S \longrightarrow \mathbb{R}_{\geq 0}$ such that $\sum\{w(x) \mid x \in S\} = 1$ and
2. an injective function $\varphi : \mathcal{P}(S) \longrightarrow (S \longrightarrow \mathbb{R})$.

The metric defined by the pair $(w, \varphi)$ is the function $d_{w,\varphi} : \mathcal{P}(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by

$$d_{w,\varphi}(X, Y) = \left( \sum_{s \in S} w(s) |\varphi(X)(s) - \varphi(Y)(s)|^q \right)^{\frac{1}{q}}$$

for $X, Y \in \mathcal{P}(S)$.

The function $w$ is extended to $\mathcal{P}(S)$ by

$$w(T) = \sum\{w(x) \mid x \in T\}.$$

Clearly, $w(\emptyset) = 0$ and $w(S) = 1$. Also, if $P$ and $Q$ are two disjoint subsets, we have $w(P \cup Q) = w(P) + w(Q)$.

We refer to both $w$ and its extension to $\mathcal{P}(S)$ as *weight functions*.

The value $\varphi(T)$ of the function $\varphi$ is itself a function $\varphi(T) : S \longrightarrow \mathbb{R}$, and each subset $T$ of $S$ defines such a distinct function. These notions are used in the next theorem.

**Theorem 10.65.** *Let $S$ be a set, $w : S \longrightarrow \mathbb{R}_{\geq 0}$ be a weight function, and $\varphi : \mathcal{P}(S) \longrightarrow (S \longrightarrow \mathbb{R})$ be an injective function.*

*If $w(x) > 0$ for every $x \in S$, then the mapping $d_{w,\varphi} : (\mathcal{P}(S))^2 \longrightarrow \mathbb{R}$ defined by*

$$d_{w,\varphi}(U,V) = \left(\sum_{x\in S} w(x)|\varphi(U)(x) - \varphi(V)(x)|^p\right)^{\frac{1}{p}} \tag{10.10}$$

*for $U, V \in \mathcal{P}(S)$ is a metric on $\mathcal{P}(S)$.*

*Proof.* It is clear that $d_{w,\varphi}(U,U) = 0$. If $d_{w,\varphi}(U,V) = 0$, then $\varphi(U)(x) = \varphi(V)(x)$ because $w(x) > 0$, for every $x \in S$. Thus, $\varphi(U) = \varphi(V)$, which implies $U = V$ due to the injectivity of $\varphi$.

The symmetry of $d_{w,\varphi}$ is immediate.

To prove the triangular inequality, we apply Minkowski's inequality. Suppose that $S = \{x_0, \ldots, x_{n-1}\}$, and let $U, V, W \in \mathcal{P}(S)$. Define the numbers

$$\begin{aligned} a_i &= (w(x_i))^{\frac{1}{p}}\varphi_U(x_i),\\ b_i &= (w(x_i))^{\frac{1}{p}}\varphi_V(x_i),\\ c_i &= (w(x_i))^{\frac{1}{p}}\varphi_W(x_i), \end{aligned}$$

for $0 \leq i \leq n-1$. Then, by Minkowski's inequality, we have

$$\left(\sum_{i=0}^{n-1} |a_i - b_i|^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=0}^{n-1} |a_i - c_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=0}^{n-1} |c_i - b_i|^p\right)^{\frac{1}{p}},$$

which amounts to the triangular inequality $d_{w,\varphi}(U,V) \leq d_{w,\varphi}(U,W) + d_{w,\varphi}(W,V)$. Thus, we may conclude that $d_{w,\varphi}$ is indeed a metric on $\mathcal{P}(S)$. □

*Example 10.66.* Let $w : S \longrightarrow [0,1]$ be a positive weight function. Define the function $\varphi$ by

$$\varphi(U)(x) = \begin{cases} \frac{1}{\sqrt{w(U)}} & \text{if } x \in U,\\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $\varphi(U) = \varphi(V)$ if and only if $U = V$, so $\varphi$ is an injective function.

Choosing $p = 2$, the metric defined in Theorem 10.65 becomes

$$d^2_{w,\varphi}(U,V) = \left(\sum_{x\in S} w(x)|\varphi(U)(x) - \varphi(V)(x)|^2\right)^{\frac{1}{2}}.$$

Suppose initially that neither $U$ nor $V$ are empty. Several cases need to be considered:

1. If $x \in U \cap V$, then

$$|\varphi(U)(x) - \varphi(V)(x)|^2 = \frac{1}{w(U)} + \frac{1}{w(V)} - \frac{2}{\sqrt{w(U)w(V)}}.$$

The total contribution of these elements of $S$ is

$$w(U \cap V)\left(\frac{1}{w(U)} + \frac{1}{w(V)} - \frac{2}{\sqrt{w(U)w(V)}}\right).$$

If $x \in U - V$, then

$$|\varphi(U)(x) - \varphi(V)(x)|^2 = \frac{1}{w(U)}$$

and the total contribution is $w(U - V)\frac{1}{w(U)}$.

2. When $x \in V - U$, then

$$|\varphi(U)(x) - \varphi(V)(x)|^2 = \frac{1}{w(V)}$$

and the total contribution is $w(V - U)\frac{1}{w(V)}$.

3. Finally, if $x \notin U \cup V$, then $|\varphi(U)(x) - \varphi(V)(x)|^2 = 0$.

Thus, we can write

$$\begin{aligned} d^2_{w,\varphi}(U,V) &= w(U \cap V)\left(\frac{1}{w(U)} + \frac{1}{w(V)} - \frac{2}{\sqrt{w(U)w(V)}}\right) \\ &\quad + w(U - V)\frac{1}{w(U)} + w(V - U)\frac{1}{w(V)} \\ &= \frac{w(U \cap V) + w(U - V)}{w(U)} + \frac{w(U \cap V) + w(V - U)}{w(V)} \\ &\quad - \frac{2w(U \cap V)}{\sqrt{w(U)w(V)}} \\ &= 2\left(1 - \frac{w(U \cap V)}{\sqrt{w(U)w(V)}}\right), \end{aligned}$$

where we used the fact that $w(U \cap V) + w(U - V) = w(U)$ and $w(U \cap V) + w(V - U) = w(V)$. Thus,

$$d_{w,\varphi}(U,V) = \sqrt{2\left(1 - \frac{w(U \cap V)}{\sqrt{w(U)w(V)}}\right)}.$$

If $U \neq \emptyset$ and $V = \emptyset$, then it is immediate that $d_{w,\varphi}(U, \emptyset) = 1$. Of course, $d_{w,\varphi}(\emptyset, \emptyset) = 0$.

Thus, the mapping $d_{w,\varphi}$ defined by

$$d_{w,\varphi}(U,V) = \begin{cases} 0 & \text{if } U = V = \emptyset, \\ 1 & \text{if } U \neq \emptyset \text{ and } V = \emptyset, \\ 1 & \text{if } U = \emptyset \text{ and } V \neq \emptyset, \\ \sqrt{2\left(1 - \frac{w(U \cap V)}{\sqrt{w(U)w(V)}}\right)} & \text{if } U \neq \emptyset \text{ and } V \neq \emptyset, \end{cases}$$

for $U, V \in \mathcal{P}(S)$ is a metric, which is known as the *Ochïai metric* on $\mathcal{P}(S)$.

*Example 10.67.* Using the same notation as in Example 10.66 for a positive weight function $w : S \longrightarrow [0,1]$, define the function $\varphi$ by

$$\varphi(U)(x) = \begin{cases} \frac{1}{w(U)} & \text{if } x \in U, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $\varphi$ is an injective function.

Suppose that $p = 2$ in Equality (10.10). If $U \neq \emptyset$ and $V \neq \emptyset$, we have the following cases:

1. If $x \in U \cap V$, then

$$|\varphi(U)(x) - \varphi(V)(x)|^2 = \frac{1}{w(U)^2} + \frac{1}{w(V)^2} - \frac{2}{w(U)w(V)}.$$

The total contribution of these elements of $S$ is

$$w(U \cap V)\left(\frac{1}{w(U)^2} + \frac{1}{w(V)^2} - \frac{2}{w(U)w(V)}\right).$$

If $x \in U - V$, then

$$|\varphi(U)(x) - \varphi(V)(x)|^2 = \frac{1}{w(U)^2}$$

and the total contribution is $w(U - V)\frac{1}{w(U)^2}$.

2. When $x \in V - U$, then

$$|\varphi(U)(x) - \varphi(V)(x)|^2 = \frac{1}{w(V)^2}$$

and the total contribution is $w(V - U)\frac{1}{w(V)^2}$.

3. Finally, if $x \notin U \cup V$, then $|\varphi(U)(x) - \varphi(V)(x)|^2 = 0$.

Summing up these contributions, we can write

$$\begin{aligned} d^2_{w,\varphi}(U,V) &= \frac{1}{w(U)} + \frac{1}{w(V)} - 2\frac{w(U \cap V)}{w(U)w(V)} \\ &= \frac{w(U) + w(V) - 2w(U \cap V)}{w(U)w(V)} \\ &= \frac{w(U \oplus V)}{w(U)w(V)}. \end{aligned}$$

If $V = \emptyset$, $d_{w,\varphi}(U, \emptyset) = \sqrt{\frac{1}{w(U)}}$; similarly, $d_{w,\varphi}(\emptyset, V) = \sqrt{\frac{1}{w(V)}}$.

We proved that the mapping $d_{w,\varphi}$ defined by

$$d_{w,\varphi}(U,V) = \begin{cases} \sqrt{\frac{w(U \oplus V)}{w(U)w(V)}} & \text{if } U \neq \emptyset \text{ and } V \neq \emptyset, \\ \sqrt{\frac{1}{w(U)}} & \text{if } U \neq \emptyset \text{ and } V = \emptyset, \\ \sqrt{\frac{1}{w(V)}} & \text{if } U = \emptyset \text{ and } V \neq \emptyset, \\ 0 & \text{if } U = V = \emptyset, \end{cases}$$

for $U, V \in \mathcal{P}(S)$, is a metric on $\mathcal{P}(S)$ known as the $\chi^2$ *metric*.

## 10.7 Metrics on Partitions

Metrics on sets of partitions of finite sets are useful in data mining because attributes induce partitions on the sets of tuples of tabular data. Thus, they help us determine interesting relationships between attributes and to use these relationships for classification, feature selection, and other applications. Also, exclusive clusterings can be regarded as partitions of the set of clustered objects, and partition metrics can be used for evaluating clusterings, a point of view presented in [97].

Let $S$ be a finite set and let $\pi$ and $\sigma$ be two partitions of $S$. The equivalence relations $\rho_\pi$ and $\rho_\sigma$ are subsets of $S \times S$ that allow us a simple way of defining a metric between $\pi$ and $\sigma$ as the relative size of the symmetric difference between the sets of pairs $\rho_\pi$ and $\rho_\sigma$,

$$\delta(\pi,\sigma) = \frac{1}{|S|^2}|\rho_\pi \oplus \rho_\sigma| = \frac{1}{|S|^2}\left(|\rho_\pi| + |\rho_\sigma| - 2|\rho_\pi \cap \rho_\sigma|\right). \tag{10.11}$$

If $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$, then there are $\sum_{i=1}^m |B_i|^2$ pairs in $\rho_\pi$, $\sum_{j=1}^n |C_j|^2$ pairs in $\rho_\sigma$, and $\sum_{i=1}^n \sum_{j=1}^n |B_i \cap C_j|^2$ pairs in $\rho_\pi \cap \rho_\sigma$. Thus, we have

$$\delta(\pi,\sigma) = \frac{1}{|S|^2}\left(\sum_{i=1}^m |B_i|^2 + \sum_{j=1}^n |C_j|^2 - 2\sum_{i=1}^n \sum_{j=1}^n |B_i \cap C_j|^2\right). \tag{10.12}$$

The same metric can be linked to a special case of a more general metric related to the notion of partition entropy.

We can now show a central result.

**Theorem 10.68.** *For every $\beta \geq 1$ the mapping $d_\beta : PART(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by*

$$d_\beta(\pi,\sigma) = \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi)$$

*for $\pi, \sigma \in PART(S)$ is a metric on $PART(S)$.*

*Proof.* A double application of Corollary 8.50 yields

$$\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi|\tau),$$
$$\mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\tau|\sigma) \geq \mathcal{H}_\beta(\tau|\pi).$$

Adding these inequalities gives

$$d_\beta(\pi, \sigma) + d_\beta(\sigma, \tau) \geq d_\beta(\pi, \tau),$$

which is the triangular inequality for $d_\beta$.

The symmetry of $d_\beta$ is obvious, and it is clear that $d_\beta(\pi, \pi) = 0$ for every $\pi \in PART(S)$.

Suppose now that $d_\beta(\pi, \sigma) = 0$. Since the values of $\beta$-conditional entropies are nonnegative, this implies $\mathcal{H}_\beta(\pi|\sigma) = \mathcal{H}_\beta(\sigma|\pi) = 0$. By Theorem 8.40, we have both $\sigma \leq \pi$ and $\pi \leq \sigma$, so $\pi = \sigma$. Thus, $d_\beta$ is a metric on $PART(S)$. □

An explicit expression of the metric between two partitions can now be obtained using the values of conditional entropies given by Equality (8.5),

$$d_\beta(\pi, \sigma) = \frac{1}{(1 - 2^{1-\beta})|S|^\beta} \left( \sum_{i=1}^{m} |B_i|^\beta + \sum_{j=1}^{n} |C_j|^\beta - 2 \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} |B_i \cap C_j|^\beta \right),$$

where $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$ are two partitions from $PART(S)$.

In the special case $\beta = 2$, we have

$$d_2(\pi, \sigma) = \frac{2}{|S|^2} \left( \sum_{i=1}^{m} |B_i|^2 + \sum_{j=1}^{n} |C_j|^2 - \sum_{i=1}^{m} \sum_{j=1}^{n} 2|B_i \cap C_j|^2 \right),$$

which implies $d_2(\pi, \sigma) = 2\delta(\pi, \sigma)$, where $\delta$ is the distance introduced by using the symmetric difference in Equality (10.12).

It is clear that $d_\beta(\pi, \omega_S) = \mathcal{H}_\beta(\pi)$ and $d_\beta(\pi, \alpha_S) = \mathcal{H}(\alpha_S|\pi)$. Another useful form of $d_\beta$ can be obtained by applying Theorem 8.41. Since $\mathcal{H}_\beta(\pi|\sigma) = \mathcal{H}_\beta(\pi \wedge \sigma) - \mathcal{H}_\beta(\sigma)$ and $\mathcal{H}_\beta(\sigma|\pi) = \mathcal{H}_\beta(\pi \wedge \sigma) - \mathcal{H}_\beta(\sigma)$, we have

$$d_\beta(\pi, \sigma) = 2\mathcal{H}_\beta(\pi \wedge \sigma) - \mathcal{H}_\beta(\pi) - \mathcal{H}_\beta(\sigma), \tag{10.13}$$

for $\pi, \sigma \in PART(S)$.

The behavior of the metric $d_\beta$ with respect to partition addition is discussed in the next statement.

**Theorem 10.69.** *Let $S$ be a finite set and $\pi$ and $\theta$ be two partitions of $S$, where $\theta = \{D_1, \ldots, D_h\}$. If $\sigma_i \in PART(D_i)$ for $1 \leq i \leq h$, then we have $\sigma_1 + \cdots + \sigma_h \leq \theta$ and*

$$d_\beta(\pi, \sigma_1 + \cdots + \sigma_h) = \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta d_\beta(\pi_{D_i}, \sigma_i) + \mathcal{H}_\beta(\theta|\pi).$$

*Proof.* This statement follows directly from Theorem 8.48. □

**Theorem 10.70.** *Let $\sigma$ and $\theta$ be two partitions in PART$(S)$ such that*

$$\theta = \{D_1, \ldots, D_h\}$$

*and $\sigma \leq \theta$. Then, we have*

$$d_\beta(\theta, \sigma) = \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta d_\beta(\omega_{D_i}, \sigma_{D_i}).$$

*Proof.* In Theorem 10.69, take $\pi = \theta$ and $\sigma_i = \sigma_{D_i}$ for $1 \leq i \leq h$. Then, it is clear that $\sigma = \sigma_1 + \cdots + \sigma_h$ and we have

$$d_\beta(\theta, \sigma) = \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta d_\beta(\omega_{D_i}, \sigma_{D_i})$$

because $\theta_{D_i} = \omega_{D_i}$ for $1 \leq i \leq h$. □

The next theorem generalizes a result from [97].

**Theorem 10.71.** *In the metric space $(PART(S), d_\beta)$, we have that*
(i) *if $\sigma \leq \pi$, then $d_\beta(\pi, \sigma) = \mathcal{H}_\beta(\sigma) - \mathcal{H}_\beta(\pi)$,*
(ii) *$d_\beta(\alpha_S, \sigma) + d_\beta(\sigma, \omega_S) = d_\beta(\alpha_S, \omega_S)$, and*
(iii) *$d_\beta(\pi, \pi \wedge \sigma) + d_\beta(\pi \wedge \sigma, \sigma) = d_\beta(\pi, \sigma)$*
*for all partitions $\pi, \sigma \in PART(S)$.*

*Furthermore, we have $d_\beta(\omega_T, \alpha_T) = \frac{1-|T|^{1-\beta}}{1-2^{1-\beta}}$ for every subset $T$ of $S$.*

*Proof.* The first three statements of the theorem follow immediately from Equality (10.13); the last part is an application of the definition of $d_\beta$. □

A generalization of a result obtained in [97] is contained in the next statement, which gives an axiomatization of the metric $d_\beta$.

**Theorem 10.72.** *Let $d : PART(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ be a function that satisfies the following conditions:*
*(D1) $d$ is symmetric; that is, $d(\pi, \sigma) = d(\sigma, \pi)$.*
*(D2) $d(\alpha_S, \sigma) + d(\sigma, \omega_S) = d(\alpha_S, \omega_S)$.*
*(D3) $d(\pi, \sigma) = d(\pi, \pi \wedge \sigma) + d(\pi \wedge \sigma, \sigma)$.*
*(D4) if $\sigma, \theta \in PART(S)$ such that $\theta = \{D_1, \ldots, D_h\}$ and $\sigma \leq \theta$, then we have*

$$d(\theta, \sigma) = \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta d(\omega_{D_i}, \sigma_{D_i}).$$

***(D5)*** *$d(\omega_T, \alpha_T) = \frac{1-|T|^{1-\beta}}{1-2^{1-\beta}}$ for every $T \subseteq S$.*

*Then,* $d = d_\beta$.

*Proof.* Choosing $\sigma = \alpha_S$ in axiom (D4) and using (D5), we can write

$$\begin{aligned} d(\alpha_S, \theta) &= \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta d(\omega_{D_i}, \alpha_{D_i}) \\ &= \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta \frac{1 - |D_i|^{1-\beta}}{1 - 2^{1-\beta}} \\ &= \frac{\sum_{i=1}^{h} |D_i|^\beta - |S|}{(1 - 2^{1-\beta})|S|^\beta}. \end{aligned}$$

From Axioms (D2) and (D5) it follows that

$$\begin{aligned} d(\theta, \omega_S) &= d(\alpha_S, \omega_S) - d(\alpha_S, \theta) \\ &= \frac{1 - |S|^{1-\beta}}{1 - 2^{1-\beta}} - \frac{\sum_{i=1}^{h} |D_i|^\beta - |S|}{(1 - 2^{1-\beta})|S|^\beta} \\ &= \frac{|S|^\beta - \sum_{i=1}^{h} |D_i|^\beta}{(1 - 2^{1-\beta})|S|^\beta}. \end{aligned}$$

Now let $\pi, \sigma \in PART(S)$, where $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$. Since $\pi \wedge \sigma \leq \pi$ and $\sigma_{B_i} = \{C_1 \cap B_i, \ldots, C_n \cap B_i\}$, an application of Axiom (D4) yields:

$$\begin{aligned} d(\pi, \pi \wedge \sigma) &= \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta d(\omega_{B_i}, (\pi \wedge \sigma)_{B_i}) \\ &= \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta d(\omega_{B_i}, \sigma_{B_i}) \\ &= \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta \frac{|B_i|^\beta - \sum_{j=1}^{n} |B_i \cap C_j|^\beta}{(1 - 2^{1-\beta})|B_i|^\beta} \\ &= \frac{1}{(1 - 2^{1-\beta})|S|^\beta} \left( \sum_{i=1}^{m} |B_i|^\beta - \sum_{j=1}^{n} \sum_{i=1}^{n} |B_i \cap C_j|^\beta \right) \end{aligned}$$

because $(\pi \wedge \sigma)_{B_i} = \sigma_{B_i}$. □

By Axiom (D1), we obtain the similar equality

$$d(\pi \wedge \sigma, \sigma) = \frac{1}{(1 - 2^{1-\beta})|S|^\beta} \left( \sum_{i=1}^{m} |B_i|^\beta - \sum_{j=1}^{n} \sum_{i=1}^{n} |B_i \cap C_j|^\beta \right),$$

which, by Axiom (D3), implies:

$$d(\pi,\sigma) = \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(\sum_{i=1}^{m}|B_i|^\beta + \sum_{j=1}^{n}|C_j|^\beta - 2\sum_{j=1}^{n}\sum_{i=1}^{n}|B_i \cap C_j|^\beta\right);$$

that is, $d(\pi,\sigma) = d_\beta(\pi,\sigma)$. □

## 10.8 Metrics on Sequences

Sequences are the objects of many data mining activities (text mining, biological applications) that require evaluation of the degree to which they are different from each other. For a fixed $n$, we can easily define a metric on the set $\mathbf{Seq}_n(S)$ as

$$d(\mathbf{x},\mathbf{y}) = |\{i \mid 0 \leq i \leq n-1, x_i \neq y_i\}|,$$

for every $\mathbf{x} = (x_0,\ldots,x_{n-1})$ and $\mathbf{y} = (y_0,\ldots,y_{n-1})$ in $\mathbf{Seq}_n(S)$. This is a rather rudimentary distance known as the *Hamming distance* on $\mathbf{Seq}_n(S)$. If we need to compare sequences of unequal length, we can use an extended metric $d'$ defined by

$$d'(\mathbf{x},\mathbf{y}) = \begin{cases} |\{i \mid 0 \leq i \leq |\mathbf{x}|-1, x_i \neq y_i\}| & \text{if } |\mathbf{x}| = |\mathbf{y}|, \\ \infty & \text{if } |\mathbf{x}| \neq |\mathbf{y}|. \end{cases}$$

The Hamming distance is not very useful in this context due to its inability to measure anything but the degree of coincidence between symbols that occur in similar position. A much more useful tool is Levenshtein's distance, introduced in [88], using certain operations on sequences.

Recall that we introduced the notion of replacement of an occurrence $(\mathbf{y}, i)$ in a sequence $\mathbf{x}$ in Definition 1.94 on page 28.

**Definition 10.73.** *Let $S$ be a set and let $\boldsymbol{x} \in \mathbf{Seq}(S)$. The* insertion of $s \in S$ in $\boldsymbol{x}$ at position $i$ *yields the sequence*

$$\mathsf{i}_{s,i}(\boldsymbol{x}) = \mathtt{replace}\,(\boldsymbol{x}, (\boldsymbol{\lambda}, i), s),$$

*where* $0 \leq i \leq |\boldsymbol{x}|$.

*The* deletion of the symbol located at position $i$ *yields the sequence:*

$$\mathsf{d}_i(\boldsymbol{x}) = \mathtt{replace}\,(\boldsymbol{x}, (\boldsymbol{x}(i), i), \boldsymbol{\lambda}),$$

*where* $0 \leq i \leq |\boldsymbol{x}| - 1$.

*The* substitution of $s \in S$ at position $i$ by $s'$ *produces the sequence*

$$\mathsf{s}_{s,i,s'}(\boldsymbol{x}) = \mathtt{replace}\,(\boldsymbol{x}, (s, i), s'),$$

*where* $0 \leq i \leq |\boldsymbol{x}| - 1$.

In Definition 10.73, we introduced three types of partial functions on the set of sequences $\mathbf{Seq}(S)$, $\mathsf{i}_{s,i}$, $\mathsf{d}_i$, and $\mathsf{s}_{s,i,s'}$, called *insertion*, *deletion*, and *substitution*, respectively. There partial functions are collectively referred to as *editing functions*. Observe that, in order to have $\mathbf{x} \in \mathrm{Dom}(\mathsf{d}_i)$, we must have $|\mathbf{x}| \geq i$.

**Definition 10.74.** *An* edit transcript *is a sequence* $(f_0, f_1, \ldots, f_{k-1})$ *of edit operations.*

*Example 10.75.* Let $S$ be the set of small letters of the Latin alphabet, $S = \{a, b, \ldots, z\}$, and let $\mathbf{x} = (m, i, c, k, e, y)$, $\mathbf{y} = (m, o, u, s, e)$. The following sequence of operations transforms $\mathbf{x}$ into $\mathbf{y}$:

| Step | Sequence | Operation |
|---|---|---|
| 0 | $(m, i, c, k, e, y)$ | $\mathsf{s}_{i,1,o}$ |
| 1 | $(m, o, c, k, e, y)$ | $\mathsf{s}_{c,2,u}$ |
| 2 | $(m, o, u, k, e, y)$ | $\mathsf{d}_3$ |
| 3 | $(m, o, u, e, y)$ | $\mathsf{d}_3$ |
| 4 | $(m, o, u, y)$ | $\mathsf{i}_{s,3}$ |
| 5 | $(m, o, u, s, y)$ | $\mathsf{s}_{y,4,e}$ |
| 6 | $(m, o, u, s, e)$ | |

The sequence

$$(\mathsf{s}_{i,1,o}, \mathsf{s}_{c,2,u}, \mathsf{d}_3, \mathsf{d}_3, \mathsf{i}_{s,3}, \mathsf{s}_{y,4,e})$$

is an edit transcript of length 6.

If $(f_0, f_1, \ldots, f_{k-1})$ is an edit transcript that transforms a sequence $\mathbf{x}$ into a sequence $\mathbf{y}$, then we have the sequences $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_k$ such that $\mathbf{z}_0 = \mathbf{x}$, $\mathbf{z}_i \in \mathrm{Dom}(f_i)$, and $f_i(\mathbf{z}_i) = \mathbf{z}_{i+1}$ for $0 \leq ik - 1$ and $\mathbf{z}_k = \mathbf{y}$. Moreover, we can write

$$\mathbf{y} = f_{k-1}(\cdots f_1(f_0(\mathbf{x})) \cdots).$$

**Theorem 10.76.** *Let* $\ell : \mathbf{Seq}(S) \times \mathbf{Seq}(S) \longrightarrow \mathbb{R}_{\geq 0}$ *be a function defined by* $\ell(\boldsymbol{x}, \boldsymbol{y}) = n$ *if* $n$ *is the length of the shortest edit transcript needed to transform* $\boldsymbol{x}$ *into* $\boldsymbol{y}$*. The function* $\ell$ *is a metric on* $\mathbf{Seq}(S)$*.*

*Proof.* It is clear that $\ell(\mathbf{x}, \mathbf{x}) = 0$ and that $\ell(\mathbf{x}, \mathbf{y}) = \ell(\mathbf{y}, \mathbf{x})$ for every $\mathbf{x}, \mathbf{y} \in \mathbf{Seq}(S)$. Observe that the triangular inequality is also satisfied because the sequence of operations that transform $\mathbf{x}$ into $\mathbf{y}$ followed by the sequence of operations that transform $\mathbf{y}$ into $\mathbf{z}$ will transform $\mathbf{x}$ into $\mathbf{z}$. Since the smallest such number of transformations is $\ell(\mathbf{x}, \mathbf{z})$, it follows that

$$\ell(\mathbf{x}, \mathbf{z}) \leq \ell(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}, \mathbf{z}).$$

This allows us to conclude that $\ell$ is a metric on $\mathbf{Seq}(S)$. □

We refer to $\ell$ as the *Levenshtein distance* between $\mathbf{x}$ and $\mathbf{y}$.

Recall that we introduced on page 27 the notation $\mathbf{x}_{i,j}$ for the infix $(x_i, \ldots, x_j)$ of a sequence $\mathbf{x} = (x_0, \ldots, x_{n-1})$.

Let $\mathbf{x} = (x_0, \ldots, x_{n-1})$ and $\mathbf{y} = (y_0, \ldots, y_{m-1})$ be two sequences and $l_{ij}(\mathbf{x}, \mathbf{y})$ be the length of a shortest edit transcript needed to transform $\mathbf{x}_{0,i}$ $\mathbf{y}_{0,j}$ for $-1 \leq i \leq |\mathbf{x}|$ and $-1 \leq j \leq |\mathbf{y}|$. In other words,

$$l_{ij} = \ell(x_{0,i}, y_{0,j}) \tag{10.14}$$

for $0 \leq i \leq n-1$ and $0 \leq j \leq m-1$, where $n = |\mathbf{x}|$ and $m = |\mathbf{y}|$.

When $i = -1$, we have $\mathbf{x}_{0,-1} = \boldsymbol{\lambda}$; similarly, when $j = -1$, $\mathbf{y}_{0,-1} = \boldsymbol{\lambda}$. Therefore, $l_{-1,j} = j$ since we need to insert $j$ elements of $S$ into $\boldsymbol{\lambda}$ to obtain $y_{0,j-1}$ and $l_{i,-1} = i$ for similar reasons.

To obtain an inductive expression of $l_{ij}$, we distinguish two cases. If $x_i = y_j$, then $l_{i,j} = l_{i-1,j-1}$; otherwise (that is, if $x_i \neq y_i$) we need to choose the edit transcript of minimal length among the following edit transcripts:

(i) the shortest edit transcript that transforms $\mathbf{x}_{0,i}$ into $\mathbf{y}_{0,j-1}$ followed by $\mathsf{i}_{x_i,j}$;
(ii) the shortest edit transcript that transforms $\mathbf{x}_{0,i-1}$ into $\mathbf{y}_{0,j}$ followed by $\mathsf{d}_i$;
(iii) the shortest edit transcript that transforms $\mathbf{x}_{0,i-1}$ into $\mathbf{y}_{0,j-1}$ followed by substitution $\mathsf{s}_{x_i,i,y_j}$ if $x_i \neq y_j$.

Therefore,

$$l_{ij} = \min\{l_{i-1,j} + 1, l_{i,j-1} + 1, l_{i-1,j-1} + \delta(i,j)\}, \tag{10.15}$$

where

$$\delta(i,j) = \begin{cases} 0 & \text{if } x_i = y_j \\ 1 & \text{otherwise.} \end{cases}$$

The numbers $l_{ij}$ can be computed using a bidimensional $(m+1) \times (n+1)$ array $L$. The rows of the array are numbered from $-1$ to $|\mathbf{x}| - 1$, while the columns are numbered from $-1$ to $|\mathbf{y}| - 1$. The component $L_{ij}$ of $L$ consists of a pair of the form $(l_{ij}, A_{ij})$, where $A_{ij}$ is a subset of the set $\{\uparrow, \leftarrow, \nwarrow\}$.

Initially, the first row of $L$ is $L_{-1,j} = (l_{-1,j}, \{\leftarrow\})$ for $-1 \leq j \leq |\mathbf{y}| - 1$; the first column of $L$ is $L_{i,-1} = (l_{i,-1}, \uparrow)$ for $-1 \leq i \leq |\mathbf{x}| - 1$.

For each of the numbers $l_{i-1,j} + 1$, $l_{i,j-1} + 1$, or $l_{i-1,j-1} + \delta(i,j)$ that equals $l_{ij}$, we include in $A_{ij}$ the symbols $\uparrow$, $\rightarrow$, or $\nwarrow$, respectively, pointing to the surrounding cells that help define $l_{ij}$. This will allow us to extract an edit transcript by following the points backward from $L_{m-1,n-1}$ to the cell $L_{-1,-1}$. Each symbol $\leftarrow$ denotes an insertion of $y_j$ into the current string, each symbol $\uparrow$ as a deletion of $x_i$ from the current string, and each diagonal edge as a match between $x_i$ and $y_j$ or as a substitution of $x_i$ by $y_j$.

*Example 10.77.* Consider the sequences $\mathbf{x} = (a, b, a, b, c, a, b, a, c)$ and $\mathbf{y} = (a, b, c, a, a, c)$.

| | | | a | b | c | a | a | c |
|---|---|---|---|---|---|---|---|---|
| | | -1 | 0 | 1 | 2 | 3 | 4 | 5 |
| | -1 | 0 | $\leftarrow 0$ | $\leftarrow 0$ | $\leftarrow 0$ | $\leftarrow 0$ | $\leftarrow 0$ | $\leftarrow 0$ |
| a | 0 | $\uparrow 0$ | $\nwarrow 0$ | $\leftarrow \nwarrow 1$ | $\leftarrow \nwarrow 2$ | $\nwarrow 2$ | $\leftarrow \nwarrow 3$ | $\leftarrow 4$ |
| b | 1 | $\uparrow 1$ | $\uparrow \nwarrow 1$ | $\nwarrow 1$ | $\leftarrow \nwarrow 2$ | $\leftarrow \uparrow \nwarrow 3$ | $\nwarrow 3$ | $\leftarrow \nwarrow 4$ |
| a | 2 | $\uparrow 2$ | $\nwarrow 1$ | $\leftarrow \uparrow \nwarrow 2$ | $\nwarrow 2$ | $\nwarrow 2$ | $\leftarrow \nwarrow 3$ | $\leftarrow \nwarrow 4$ |
| b | 3 | $\uparrow 3$ | $\uparrow 2$ | $\nwarrow 1$ | $\uparrow \nwarrow 3$ | $\uparrow \nwarrow 3$ | $\nwarrow 3$ | $\uparrow \nwarrow 4$ |
| c | 4 | $\uparrow 4$ | $\uparrow 3$ | $\uparrow 2$ | $\nwarrow 1$ | $\uparrow \nwarrow 4$ | $\uparrow \nwarrow 4$ | $\nwarrow 3$ |
| a | 5 | $\uparrow 5$ | $\uparrow \nwarrow 4$ | $\uparrow 3$ | $\uparrow 2$ | $\nwarrow 1$ | $\leftarrow 2$ | $\leftarrow 3$ |
| b | 6 | $\uparrow 6$ | $\uparrow 5$ | $\uparrow \nwarrow 4$ | $\uparrow 3$ | $\uparrow 2$ | $\nwarrow 2$ | $\leftarrow \nwarrow 3$ |
| a | 7 | $\uparrow 7$ | $\uparrow \nwarrow 6$ | $\uparrow 5$ | $\uparrow 4$ | $\uparrow \nwarrow 3$ | $\uparrow \nwarrow 3$ | $\nwarrow 3$ |
| c | 8 | $\uparrow 8$ | $\uparrow 7$ | $\uparrow 6$ | $\uparrow \nwarrow 5$ | $\uparrow 4$ | $\uparrow \nwarrow 4$ | $\nwarrow 3$ |

The content of $L_{8,5}$ shows that $\ell(\mathbf{x}, \mathbf{y}) = 3$. Following the path

| $L_{8,5}$ | $L_{7,4}$ | $L_{6,3}$ | $L_{5,3}$ | $L_{4,2}$ | $L_{3,1}$ | $L_{2,0}$ | $L_{1,0}$ | $L_{0,0}$ |
|---|---|---|---|---|---|---|---|---|
| $\nwarrow 3$ | $\uparrow \nwarrow 3$ | $\uparrow 2$ | $\nwarrow 1$ | $\nwarrow 1$ | $\nwarrow 1$ | $\uparrow 1$ | $\uparrow 0$ | 0 |

that leads from $L_{8,5}$ to $L_{0,0}$, we obtain the following edit transcript:

| Step | Sequence | Operation | Remark |
|---|---|---|---|
| 0 | $(a, b, a, b, c, a, b, a, c)$ | | match $x_8 = y_5 = c$ |
| 1 | $(a, b, a, b, c, a, b, a, c)$ | | match $x_7 = y_4 = a$ |
| 2 | $(a, b, a, b, c, a, a, c)$ | $\mathsf{d}_6$ | |
| 3 | $(a, b, a, b, c, a, a, c)$ | | match $x_5 = y_3 = a$ |
| 4 | $(a, b, a, b, c, a, a, c)$ | | match $x_4 = y_2 = c$ |
| 5 | $(a, b, a, b, c, a, a, c)$ | | match $x_3 = y_1 = b$ |
| 6 | $(a, b, a, b, c, a, a, c)$ | | match $x_2 = y_0 = a$ |
| 7 | $(a, a, b, c, a, a, c)$ | $\mathsf{d}_1$ | |
| 8 | $(a, b, c, a, a, c)$ | $\mathsf{d}_0$ | |

The notion of an edit distance can be generalized by introducing costs for the edit functions.

**Definition 10.78.** *A* cost scheme *is a triple* $(c_i, c_d, c_s) \in \hat{\mathbb{R}}_{\geq 0}$*, where the components* $c_i, c_d$*, and* $c_s$ *are referred to as the costs of an insertion, deletion, and substitution, respectively.*

The *cost of an edit transcript* $\mathbf{t} = (f_0, f_1, \ldots, f_{k-1})$ according to the cost scheme $(c_i, c_d, c_s)$ is $n_i c_i + n_d c_d + n_s c_s$, where $n_i, n_d$, and $n_s$ are the number of insertions, deletions, and substitutions that occur in $\mathbf{t}$.

When $c_i = c_d = c_s = 1$, the cost of $\mathbf{t}$ equals the length of $\mathbf{t}$, and finding the Levenshtein distance between two strings $\mathbf{x}, \mathbf{y}$ can now be seen as determining the length of the shortest editing transcript that transforms $\mathbf{x}$ into $\mathbf{y}$ using the cost schema $(1, 1, 1)$. It is interesting to remark that, for any cost schema, the minimal cost of a transcript that transforms $\mathbf{x}$ into $\mathbf{y}$ remains an extended metric on $\mathbf{Seq}(S)$. This can be shown using an argument that is similar to the one we used in Theorem 10.76.

Note that a substitution can always be replaced by a deletion followed by an insertion. Therefore, for a cost scheme $(1, 1, \infty)$, the edit transcript of minimal cost will include only insertions and deletions. Similarly, if $c_s = 1$ and $c_i = c_d = \infty$, then the edit transcript will contain only substitutions if the two sequences have equal lengths and the distance between strings will be reduced to the Hamming distance.

The recurrence (10.15) that allowed us to compute the length of the shortest edit transcript is now replaced by a recurrence that allows us to compute the least cost $C_{ij}$ of transforming the prefix $\mathbf{x}_{0,i}$ into $\mathbf{y}_{0,j}$:

$$C_{ij} = \min\{C_{i-1,j} + c_i, l_{i,j-1} + c_d, l_{i-1,j-1} + \delta(i,j)c_s\}. \tag{10.16}$$

The computation of the edit distance using the cost scheme $(c_i, c_d, c_s)$ now proceeds in a tabular manner similar to the one used for computing the length of the shortest edit transcript.

## 10.9 Searches in Metric Spaces

Searches that seek to identify objects that reside in the proximity of other objects are especially important in data mining, where the keys or the ranges of objects of interest are usually unknown. This type of search is also significant for multimedia databases, where classical, exact searches are often meaningless. For example, querying an image database to find images that contain a sunrise is usually done by providing an example image and then, identifying those images that are similar to the example. The natural framework for executing such searches is provided by metric spaces or, more generally, by dissimilarity spaces [27], and we will examine the usefulness of metric properties for efficient searching algorithms. We show how various metric properties benefit the design of searching algorithms.

Starting from a finite collection of members of $S$, $T \subseteq S$, and a *query object* $q$, we consider two types of searching problems:

(i) *range queries* that seek to compute the set $B(q, r) \cap T$, for some positive number $r$, and

(ii) *k-nearest-neighbor queries* that seek to compute a set $N_k$ such that $N_k \subseteq T$, $|N_k| = k$ and for every $x \in N_k$ and $y \in T - N_k$, we have $d(x, q) \leq d(y, q)$.

In the case of $k$-nearest-neighbor queries the set $N_k$ is not uniquely identified because of the ties that may exist. For $k = 1$, we obtain the *nearest-neighbor queries*.

The triangular inequality that is satisfied by every metric plays an essential role in reducing the amount of computation required by proximity queries. Suppose that we select an element $p$ of $S$ (referred to as a *pivot*) and we compute prior to executing any proximity searches the set of distances $\{d(p, x) \mid x \in S\}$. If we need to compute a range query $B(q, r) \cap T$, then by

the triangular inequality, we have $d(q,x) \geq |d(p,q) - d(p,x)|$. Since the distances $d(p,q)$ and $d(p,x)$ have been already computed, we can exclude from the search all elements $x$ such that $|d(p,q) - d(p,x)| > r$.

The triangular inequality also ensures that the results of the search are plausible. Indeed, it is expected that if both $x$ and $y$ are in the proximity of $q$, then a certain degree of similarity exists between $x$ and $y$. This is ensured by the triangular inequality that requires $d(x,y) \leq d(x,q) + d(q,y)$.

To execute any of these searches, we need to examine the entire collection of objects $C$ unless specialized data structures called *indexes* are prepared in advance.

One of the earliest types of indexes is the *Burkhard-Keller tree* (see [25]), which can be used for metric spaces where the distance is discrete; that is, the range of the distance function is limited to a finite set. To simplify our presentation, we assume that $\mathrm{Ran}(d) = \{0, 1, \ldots, k\}$, where $k \in \mathbb{N}$.

**Algorithm 10.79 (Construction of the Burkhard-Keller Tree)**
**Input:** a collection of elements $C$ of a metric space $(S,d)$, where $\mathrm{Ran}(d) = \{0,1,\ldots,k\}$.
**Output:** a tree $\mathcal{T}_C$ whose nodes are labeled by objects of $C$.
**Method:**
**if** $|C| = 1$, return a single-vertex tree whose root is labeled $p$;
**else**
select randomly an object $p \in C$ to serve as root of $\mathcal{T}_C$;
partition $C$ into the sets $C_1, \ldots, C_k$ defined by
$C_i = \{o \in C \mid d(o,p) = i\}$ for $1 \leq i \leq k$;
construct the trees corresponding to $C_{l_0}, \ldots, C_{l_{m-1}}$, which are the nonempty sets among $C_1, \ldots, C_k$;
connect the trees $\mathcal{T}_{C_{l_0}}, \ldots, \mathcal{T}_{C_{l_{m-1}}}$ to $p$;
return $\mathcal{T}_C$

*Example 10.80.* Consider the collection of points $\{o_1, \ldots, o_{16}\}$ in $\mathbb{R}^2$ shown in Figure 10.13. Starting from their Euclidean distance $d_2(o_i, o_j)$, we construct the discrete distance $d$ as in Exercise 17), namely, we define $d(o_i, o_j) = \lceil d_2(o_i, o_j) \rceil$ for $1 \leq i, j \leq 16$.

The Manhattan distances $d_1(o_i, o_j)$ are given by the following matrix and we shall use this distance to construct the Burkhard-Keller tree.

$$\mathbf{D} = \begin{pmatrix}
0 & 1 & 2 & 3 & 1 & 2 & 3 & 4 & 2 & 3 & 4 & 5 & 3 & 4 & 5 & 6 \\
1 & 0 & 1 & 2 & 2 & 1 & 2 & 3 & 3 & 3 & 3 & 4 & 4 & 3 & 4 & 5 \\
2 & 1 & 0 & 1 & 3 & 2 & 1 & 2 & 4 & 3 & 2 & 3 & 5 & 4 & 3 & 4 \\
3 & 2 & 1 & 0 & 4 & 3 & 2 & 1 & 9 & 4 & 3 & 2 & 6 & 5 & 4 & 3 \\
1 & 2 & 3 & 4 & 0 & 1 & 2 & 3 & 1 & 2 & 3 & 4 & 2 & 3 & 4 & 5 \\
2 & 1 & 2 & 3 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 3 & 3 & 2 & 3 & 4 \\
3 & 2 & 1 & 2 & 2 & 1 & 0 & 1 & 3 & 2 & 1 & 2 & 4 & 3 & 2 & 3 \\
4 & 2 & 2 & 1 & 3 & 2 & 1 & 0 & 4 & 3 & 2 & 1 & 5 & 4 & 3 & 2 \\
2 & 3 & 4 & 5 & 1 & 2 & 3 & 4 & 0 & 1 & 2 & 3 & 1 & 2 & 3 & 4 \\
3 & 2 & 3 & 4 & 2 & 1 & 2 & 3 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 3 \\
4 & 3 & 2 & 3 & 3 & 2 & 1 & 2 & 2 & 1 & 0 & 1 & 3 & 2 & 1 & 2 \\
5 & 4 & 3 & 2 & 4 & 3 & 2 & 1 & 3 & 2 & 1 & 0 & 4 & 3 & 2 & 1 \\
3 & 4 & 5 & 6 & 2 & 3 & 4 & 5 & 1 & 2 & 3 & 4 & 0 & 1 & 2 & 3 \\
4 & 3 & 4 & 5 & 3 & 2 & 3 & 4 & 2 & 1 & 2 & 3 & 1 & 0 & 1 & 2 \\
5 & 4 & 3 & 4 & 4 & 3 & 2 & 3 & 3 & 2 & 1 & 2 & 2 & 1 & 0 & 1 \\
6 & 5 & 4 & 3 & 5 & 4 & 3 & 2 & 4 & 3 & 2 & 1 & 3 & 2 & 1 & 0
\end{pmatrix}.$$

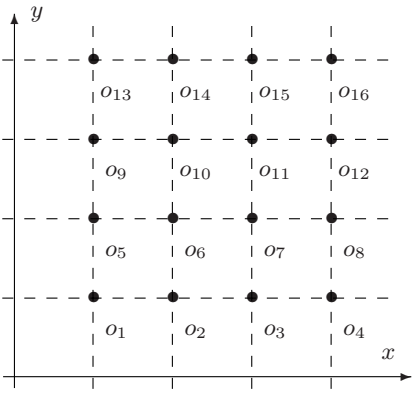



**Fig. 10.13.** Set of 16 points in $\mathbb{R}^2$.

We begin by selecting $o_6$ as the first pivot. Then, we create trees for the sets

$$C_1 = \{o_2, o_5, o_7, o_{10}\},$$
$$C_2 = \{o_1, o_3, o_8, o_{11}, o_{14}\},$$
$$C_3 = \{o_4, o_{12}, o_{13}, o_{15}\},$$
$$C_4 = \{o_{16}\}.$$

Choose $o_7$, $o_8$, and $o_{13}$ as pivots for the sets $C_1, C_2$, and $C_3$, respectively. Note that $\mathcal{T}_{C_4}$ is completed because it consists of one vertex. Assuming certain choices of pivots, the construction results in a tree shown in Figure 10.14.

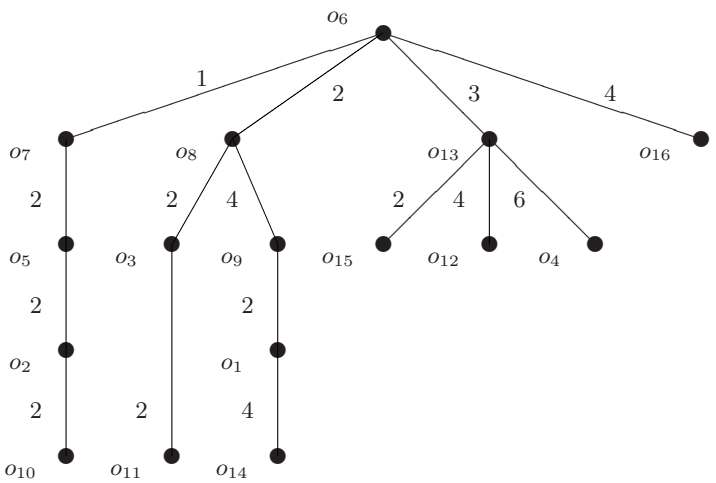



**Fig. 10.14.** Burkhard-Keller tree.

Burkhard-Keller trees can be used for range queries that seek to compute sets of the form $O_{q,r,C} = B(q,r) \cap C$, where $d$ is a discrete metric. In other words, we seek to locate all objects $o$ of $C$ such that $d(q,o) \leq r$.

By Exercise 1 we have $|d(p,o) - d(p,q)| \leq d(q,o) \leq r$, where $p$ is the pivot that labels the root of the tree. This implies $d(p,q) - r \leq d(p,o) \leq d(p,q) + r$, so we need to visit recursively the trees $\mathcal{T}_{C_i}$ where $d(p,q) - r \leq i \leq d(p,q) + r$.

**Algorithm 10.81 (Searching for Burkhard-Keller Trees)**
**Input:** a collection of elements $C$ of a metric space $(S,d)$, a query object $q$ and a radius $r$;
**Output:** the set $O(q,r,C) = B(q,r) \cap C$;
**Method:**
$O(q,r,C) = \emptyset$;
**if** $d(p,q) \leq r$ **then** $O(q,r,C) = O(q,r,C) \cup \{p\}$;
**compute** $I = \{i \mid 1 \leq i \leq k, d(p,q) - r \leq i \leq d(p,q) + r\}$;
**compute** $\bigcup_{i \in I} O(q,r,C_i)$;
$O(q,r,C) = O(q,r,C) \cup \bigcup_{i \in I} O(q,r,C_i)$;
**return** $O(q,r,C)$

*Example 10.82.* To solve the query $B(o_{11},1) \cap C$, where $C$ is the collection of objects introduced in Example 10.80, we begin by observing that $d(o_{11},o_6) = 2$, so the pivot itself does not belong to $O(o_{11},1,C)$. The set $I$ in this case is $I = \{1,2,3\}$.

We need to execute three recursive calls, namely $O_{o_{11},1,C_1}$, $O_{o_{11},1,C_2}$, and $O_{o_{11},1,C_3}$.

For the set $C_1$ having the pivot $o_7$, we have $o_7 \in O(o_{11},1,C_1)$ because $d(o_7,o_{11}) = 1$. Thus, $O(o_{11},1,C_1)$ is initialized to $\{o_7\}$ and the search proceeds with the set $C_{1,2}$, which consists of the objects $o_5, o_2, o_{10}$ located at distance 2 from the pivot $o_7$.

Since $d(o_5,o_{11}) = 3$, $o_5$ does not belong to the result. The set $C_{1,2,2}$ consists of $\{o_2,o_{10}\}$. Choosing $o_2$ as the pivot, we can exclude it from the result because $d(o_2,o_{11}) = 3$. Finally, the set $C_{1,2,2,2}$ consists of $\{o_{10}\}$ and $d(o_{10},o_{11}) = 1$. Thus,

$$O(o_{11},1,C_{1,2,2,2}) = O(o_{11},1,C_{1,2,2}) = O(o_{11},1,C_{1,2}) = \{o_{10}\},$$

so $O(o_{11},1,C_1) = \{o_7,o_{10}\}$.

Similarly, we have $O(o_{11},1,C_2) = o_{11}$ and $O(o_{11},1,C_3) = \{o_{12},o_{15}\}$. The result of the query is $O_{o_{11},1,C} = \{o_7,o_{10},o_{11},o_{12},o_{15}\}$.

Orchard's algorithm [101] aims to solve the nearest-neighbor problem and proceeds as follows.

**Algorithm 10.83 (Orchard's Algorithm)**
**Input:** a finite metric space $(S,d)$ and a query $q \in S$.

**Output:** the member of $S$ that is closest to the query $q$.
**Method:** for each $w \in S$ establish a list $L_w$ of elements of $S$
in increasing order of their distance to $w$;
(preprocessing phase)
select an initial candidate $c$;
**repeat**
compute $d(c, q)$;
scan $L_c$ until a site $s$ closer to $q$ is found;
$c = s$;
**until** ($L_c$ is completely traversed or
$s$ is found in $L_c$ such that $d(c, s) > 2d(c, q)$);
return $c$

Since $L_c$ lists all elements of the space in increasing order of their distance to $c$, observe that if the scanning of a list $L_c$ is completed without finding an element that is closer to $q$, then it is clear that $p$ is one of the elements that is closest to $q$ and the algorithm halts. Let $s$ be the first element in the current list $L_c$ such that $d(c, s) > 2d(c, q)$ (if such elements exist at all). Then, none of the previous elements on the list is closer to $q$ than $s$ since otherwise we would not have reached $s$ in $L_c$. By Exercise 5, (with $k = 2$) we have $d(q, s) > d(c, q)$, so $c$ is still the closest element to $q$ on this list. If $z$ is an element of $L_c$ situated past $s$, it follows that $d(z, c) \geq d(s, c)$ because $L_c$ is arranged in increasing order of the distances to $c$, so $d(c, z) > 2d(c, q)$, which ensures that $z$ is more distant from $q$ than $c$. So, in all cases where $c$ is closest to $q$, the algorithm works correctly.

The preprocessing phase requires an amount of space that grows as $\Theta(n^2)$ with $n$, and this limits the usefulness of the algorithm to rather small sets of objects.

An alternative algorithm known as the *annulus algorithm*, proposed in [69], allows reduction of the volume of preprocessing space to $\Theta(n)$.

Suppose that the finite metric space $(S, d)$ consists of $n$ objects. The preprocessing phase consists of selecting a pivot $o$ and constructing a list of objects $L = (o_1, o_2, \ldots, o_n)$ such that $d(o, o_1) \leq d(o, o_2) \leq \cdots \leq d(o, o_n)$. Without loss of generality, we may assume that $o = o_1$.

Suppose that $u$ is closer to the query $q$ than $v$; that is, $d(q, u) \leq d(q, v)$. Then we have

$$|d(u, o - d(q, o)| \leq d(u, q) \leq d(q, v)$$

by Exercise 1, which implies

$$d(q, o) - d(q, v) \leq d(u, o) \leq d(q, p) + d(q, v).$$

Thus, $u$ is located in an annulus centered in $o$ that contains the query point $q$ and is delimited by two spheres, $B(o, d(q, o) - d(q, v))$ and $B(o, d(q, p) + d(q, v))$.

**Algorithm 10.84 (Annulus algorithm)**
**Input:** a finite metric space $(S, d)$ and a query $q \in S$.
**Output:** the member of $S$ that is closest to the query $q$.
**Method:** select a pivot object $o$;
establish a list $L$ of elements of $S$
in increasing order of their distances to $o$;
(preprocessing phase)
select an initial candidate $v$;
compute the set of $U_v$ that consists of those $u$ such that
$d(q, o) - d(q, v) \leq d(u, o) \leq d(q, p) + d(q, v)$;
scan $U_v$ for an object $w$ closer to $q$;
if such a vector exists then
replace $v$ by $w$, recompute $U_v$ and resume scan;
otherwise, output $v$

The advantage of the annulus algorithm over Orchard's algorithm consists of the linear amount of space required for the preprocessing phase. Implementations of these algorithms and performance issues are discussed in detail in [149].

A general algorithm for the nearest-neighbor search in dissimilarity spaces was formulated in [50]. The algorithm uses the notion of *basis for a subset of a dissimilarity space.*

**Definition 10.85.** *Let $(S, d)$ be a dissimilarity space and let $(\alpha, \beta)$ be a pair of numbers such that $\alpha \geq \beta > 0$. A* basis *at level $(\alpha, \beta)$ for a subset $H$, $H \subseteq S$ is a finite set of points $\{z_1, \ldots, z_k\} \subseteq S$ if for any $x, y \in H$ we have*

$$\alpha d(x, y) \geq \max\{|d(x, z_i) - d(y, z_i)| \mid 1 \leq i \leq k\} \geq \beta d(x, y).$$

*A dissimilarity space is* $k$-dimensional *if there exist $\alpha, \beta$, and $k$ depending only on $(S, d)$ such that for any bounded subset $H$ of $S$, there are $k$ points of $S$ that form a basis at level $(\alpha, \beta)$ for $H$.*

*Example 10.86.* Consider the metric space $(\mathbb{R}^n, d_2)$, and let $H$ be a bounded subset of $\mathbb{R}^n$. A basis at level $(1, 0.5)$ for $H$ can be formed by the $n + 1$ vertices of a sufficiently large $n$-dimensional simplex $S_n$ that contains $H$. Indeed, observe that $d_2(\mathbf{x}, \mathbf{y}) \geq |d_2(\mathbf{x}, \mathbf{z}_i) - d_2(\mathbf{y}, \mathbf{z}_i)|$ for $1 \leq i \leq n + 1$ by Exercise 1, which shows that the first condition of Definition 10.85 is satisfied.

On the other hand, if the $n + 1$ points of the $n$-dimensional simplex are located at sufficient distances from the points of the set $H$, then there exists at least one vertex $\mathbf{z}_i$ such that $|d_2(\mathbf{u}, \mathbf{z}_i) - d_2(\mathbf{v}, \mathbf{z}_i)| \geq 0.5 d_2(\mathbf{u}, \mathbf{v})$; that is, $\max\{|d_2(\mathbf{u}, \mathbf{z}_i) - d_2(\mathbf{v}, \mathbf{z}_i)| \mid 1 \leq i \leq k\} \geq 0.5 d(\mathbf{u}, \mathbf{v})$. Indeed, let $h_i$ be the distance from $\mathbf{z}_i$ to the line determined by $\mathbf{u}$ and $\mathbf{v}$, and let $\mathbf{w}_i$ be the projection of $\mathbf{z}_i$ on this line (see Figure 10.15). We discuss here only the case where $\mathbf{w}_i$ is located outside the segment $(\mathbf{u}, \mathbf{v})$. Let $k_i = \min\{d_2(\mathbf{u}, \mathbf{w}_i), d_2(\mathbf{v}, \mathbf{w}_i)\}$.

To satisfy the condition

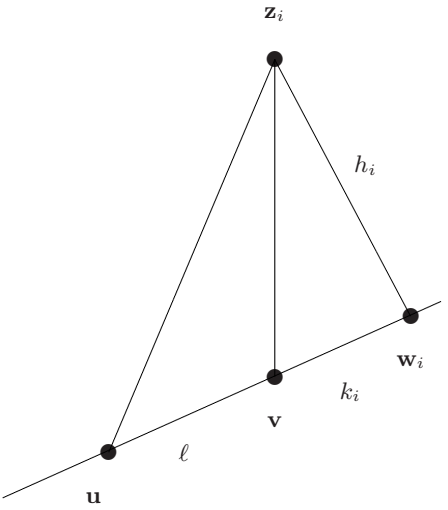



**Fig. 10.15.** Point of the basis for a set $H$ in $\mathbb{R}^n$.

$$|d_2(\mathbf{u}, \mathbf{z}_i) - d_2(\mathbf{v}, \mathbf{z}_i)| \geq \frac{\ell}{2}$$

or the equivalent equality

$$\left| \sqrt{h_i^2 + (\ell + k_i)^2} - \sqrt{h_i^2 + k_i^2} \right| \geq \frac{\ell}{2},$$

it suffices to have

$$\ell + 2k_i \geq \sqrt{h_i^2 + (\ell + k_i)^2}.$$

This inequality is satisfied if $k_i \geq \frac{1}{3}\left(\sqrt{\ell + 4h_i^2} - \ell\right)$. Thus, if $\mathbf{z}_i$ is chosen appropriately, the set $\mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{z}_{n+1}$ is a $(1, 0.5)$ basis for $H$.

**Algorithm 10.87 (Faragó-Linder-Lugosi Algorithm)**
**Input:** a collection $H = \{x_1, \ldots, x_n\}$ of a dissimilarity space $(S, d)$;
an $(\alpha, \beta)$-basis $z_1, \ldots, z_k$ for $H$, and a query $x \in S$;
**Output:** the member of $H$ that is closest to the query $x$;
**Method:** compute and store all dissimilarities $d(x_i, z_j)$ for
$1 \leq i \leq n$ and $1 \leq j \leq k$ (preprocessing phase);
$\mathcal{I} = \{x_1, \ldots, x_n\}$;
$\gamma(x_i) = \max_{1 \leq j \leq k} |d(x_i, z_j) - d(x, z_j)|$ for $1 \leq i \leq n$;
$t_0 = \min_{1 \leq i \leq n} \gamma(x_i)$;
delete all points $x_i$ from $\mathcal{I}$ such that $\gamma(x_i) > \frac{\alpha}{\beta} t_0$;
find the nearest neighbor of $x$ in the remaining part of $\mathcal{I}$ by
exhaustive search and output $x_{nn} = \arg\min_{1 \leq i \leq n} \gamma(x_i)$.

If a tie occurs in the last step of the algorithm, then an arbitrary element is chosen among the remaining elements of $\mathcal{I}$ that minimize $\gamma(x_i)$.

The first phase of the algorithm is designated as the preprocessing phase because it is independent of the query $x$ and can be executed only once for the data set $H$ and its base. Its time requirement is $O(nk)$.

To prove the correctness of the algorithm, we need to show that if an element of $H$ is deleted from $\mathcal{I}$, then it is never the nearest neighbor $x_{nn}$ of $x$. Suppose that the nearest neighbor $x_{nn}$ were removed. This would imply $\gamma(x_{nn}) > \frac{\alpha}{\beta} t_0$ or, equivalently,

$$\frac{1}{\alpha}\gamma(x_{nn}) > \frac{1}{\beta} \min_{1 \leq i \leq n} \gamma(x_i).$$

Since $\{z_1, \ldots, z_k\}$ is a basis for the set $H$, we have

$$d(x, x_{nn}) \geq \frac{1}{\alpha} \max_{1 \leq j \leq k} |d(x, z_j) - d(x_{nn}, z_j)| \geq \frac{1}{\alpha}\gamma(x_{nn})$$

and

$$\frac{1}{\beta} \min_{1 \leq i \leq n} \gamma(x_i) \geq \min_{1 \leq i \leq n} d(x, x_i),$$

which implies $d(x, x_{nn}) > \min_{1 \leq i \leq n} d(x, x_i)$. This contradicts the definition of $x_{nn}$, so it is indeed impossible to remove $x_{nn}$. Thus, in the worst case, the algorithm performs $n$ dissimilarity calculations.

Next, we present a unifying model of searching in metric spaces introduced in [27] that fits several algorithms used for proximity searches. The model relates equivalence relations (or partitions) to indexing schemes.

**Definition 10.88.** *The index defined by the equivalence $\rho$ on the set $S$ is the surjection $I_\rho : S \longrightarrow S/\rho$ defined by $I_\rho(x) = [x]$, where $[x]$ is the equivalence class of $x$ in the quotient set $S/\rho$.*

A general searching strategy can be applied in the presence of an index and involves two phases:

(i) identify the equivalence classes that contain the answers to the search, and

(ii) exhaustively search the equivalence classes identified in the first phase.

The cost of the first phase is the *internal complexity* of the search, while the cost of the second phase is the *external complexity*.

If $\rho_1, \rho_2 \in EQS(S)$ and $\rho_1 \leq \rho_2$, then $|S/\rho_1| \geq |S/\rho_2|$. Therefore, the internal complexity of the search involving $I_{\rho_1}$ is larger than the internal complexity of the search involving $I_{\rho_2}$ since we have to search more classes, while the external complexity of the search involving $I_{\rho_1}$ is smaller than the external complexity of the search involving $I_{\rho_2}$ since the classes that need to be exhaustively searched form a smaller set.

Let $(S, d)$ be a metric space and let $\rho$ be an equivalence on the set $S$. The metric $d$ generates a mapping $\delta_{d,\rho} : (S/\rho)^2 \longrightarrow \mathbb{R}_{\geq 0}$ on the quotient set $S/\rho$,

where $\delta_{d,\rho}([x],[y]) = \inf\{d(u,v) \mid u \in [x] \text{ and } v \in [y]\}$. We will refer to $\delta_{d,\rho}$ as the *pseudo-distance generated by d and ρ*. It is clear that $\delta_{d,\rho}([x],[y]) \leq d(x,y)$, for $x, y \in S$, but $\delta_{d,\rho}$ is not a metric because if fails to satisfy the triangular inequality in general.

If a range query $B(q,r) \cap T = \{y \in T \mid d(q,y) \leq r\}$ must be executed we can transfer this query on the quotient set $S/\rho$ (which is typically a smaller set than $S$) as the range query $B([q],r) \cap \{[t] \mid t \in T\}$. Note that if $y \in B(q,r)$, then $d(q,y) \leq r$, so $\delta_{d,\rho}([q],[y]) \leq d(q,y) \leq r$. This setting allows us to reduce the search of the entire set $T$ to the search of the set of equivalence classes $\{[y] \mid y \in T, \delta_{d,\rho}([q],[y]) \leq r\}$.

Since $\delta_{d,\rho}$ is not a metric, it is not possible to reduce the internal complexity of the algorithm. In such cases, a solution is to determine a metric $e$ on the quotient set $S/\rho$ such that $e([x],[y]) \leq \delta_{d,\rho}([x],[y])$ for every $[x],[y] \in S/\rho$. If this is feasible, then we can search the quotient space for classes $[y]$ such that $e([q],[y]) \leq r$ using the properties of the metric $e$.

Let $\rho_1, \rho_2 \in EQS(S)$ be two equivalences on $S$ such that $\rho_1 \leq \rho_2$. Denote by $[z]_i$ the equivalence class of $z$ relative to the equivalence $\rho_i$ for $i = 1, 2$.

If $\rho_1 \leq \rho_2$, then $[z]_1 \subseteq [z]_2$ for every $z \in S$ and therefore

$$\begin{aligned}\delta_{d,\rho_1}([x]_1,[y]_1) &= \inf\{d(u,v) \mid u \in [x]_1 \text{ and } v \in [y]_1\}\\ &\geq \inf\{d(u,v) \mid u \in [x]_2 \text{ and } v \in [y]_2\}\\ &= \delta_{d,\rho_2}([x]_2,[y]_2).\end{aligned}$$

Thus, $\delta_{d,\rho_1}([x]_1,[y]_1) \geq \delta_{d,\rho_2}([x]_2,[y]_2)$ for every $x, y \in S$, and this implies

$$\{[y]_1 \mid \delta_{d,\rho_1}([q],[y]) \leq r\} \subseteq \{[y]_2 \mid \delta_{d,\rho_2}([q],[y]) \leq r\},$$

confirming that the external complexity of the indexing algorithm based on $\rho_2$ is greater than the same complexity for the indexing algorithm based on $\rho_1$.

*Example 10.89.* Let $(S,d)$ be a metric space, $p \in S$, and let $\mathbf{r}$ be a sequence of positive real numbers $\mathbf{r} = (r_0, r_1, \ldots, r_{n-1})$ such that $r_0 < r_1 < \ldots < r_{n-1}$. Define the collection of sets $\mathcal{E} = \{E_0, E_1, \ldots, E_n\}$ by $E_0 = \{x \in S \mid d(p,x) < r_0\}$, $E_i = \{x \in S \mid r_{i-1} \leq d(p,x) < r_i\}$ for $1 \leq i \leq n-1$, and $E_n = \{x \in S \mid r_{n-1} \leq d(p,x)\}$. The subcollection of $\mathcal{E}$ that consists of nonempty sets is a partition of $S$ denoted by $\pi_{\mathbf{r}}$. Denote the corresponding equivalence relation by $\rho_{\mathbf{r}}$.

If $E_i \neq \emptyset$, then $E_i$ is an equivalence class of $\rho_{\mathbf{r}}$ that can be imagined as a circular ring around $p$. Then, if $i < j$ and $E_i$ and $E_j$ are equivalence classes, we have $\delta_{d,\rho}(E_i, E_j) > r_{j-1} - r_i$.

*Example 10.90.* Let $(S,d)$ be a metric space and $p$ be a member of $S$. Define the equivalence

$$\rho_p = \{(x,y) \in S \times S \mid d(p,x) = d(p,y)\}.$$

We have $\delta_{d,\rho}([x],[y]) = |d(p,x) - d(p,y)|$. It is easy to see that the pseudo-distance $\delta_{d,\rho}$ is actually a distance on the quotient set $S/\rho$.

## Exercises and Supplements

1. Let $(S,d)$ be a metric space. Prove that $d(x,y) \geq |d(x,z) - d(y,z)|$ for all $x,y,z \in S$.
2. Let $(S,\mathcal{E})$ be a measurable space and let $m : \mathcal{E} \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ be a measure. Prove that the $d_m$ defined by $d_m(U,V) = m(U \oplus V)$ is a semimetric on $\mathcal{E}$.
3. Let $B = \{x_1, \ldots, x_n\}$ be a finite subset of a metric space $(S,d)$. Prove that
$$(n-1)\sum_{i=1}^{n} d(x,x_i) \geq \sum\{d(x_i,x_j) \mid 1 \leq i < j \leq n\}$$
for every $x \in S$.
Explain why this inequality can be seen as a generalization of the triangular inequality.
4. Let $(S,d)$ be a metric space and let $T$ be a finite subset of $S$. Define the mapping $D_T : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ by
$$D_T(x,y) = \max\{|d(t,x) - d(t,y)| \mid t \in P\}$$
for $x,y \in S$.
   a) Prove that $D_T$ is a semimetric on $S$ and that $d(x,y) \geq D_T(x,y)$ for $x,y \in S$.
   b) Prove that if $T \subseteq T'$, then $D_T(x,y) \leq D_{T'}(x,y)$ for every $x,y \in S$.
5. Let $(S,d)$ be a metric space and let $p,q,x \in S$. Prove that $d(p,x) > kd(p,q)$ for some $k > 1$, then $d(q,x) > (k-1)d(p,q)$.
6. Let $(S,d)$ be a metric space and let $x,y \in S$. Prove that if $r$ is a positive number and $y \in C_d(x, \frac{r}{2})$, then $C_d(x, \frac{r}{2}) \subseteq C_d(y,r)$.
7. Let $(S,d)$ be a metric space and $p \in S$. Define the function $d_u : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ by
$$d_u(x,y) = \begin{cases} 0 & \text{if } x = y, \\ d(x,u) + d(u,y) & \text{otherwise,} \end{cases}$$
for $x,y \in S$. Prove that $d$ is a metric on $S$.
8. Let $f : L \longrightarrow \mathbb{R}_{\geq 0}$ be a real-valued, nonnegative function, where $\mathfrak{L} = (L, \{\wedge, \vee\})$ is a lattice. Define the mapping $d : L^2 \longrightarrow \mathbb{R}_{\geq 0}$ as $d(x,y) = 2f(x \wedge y) - f(x) - f(y)$ for $x,y \in L$. Prove that $d$ is a semimetric on $L$ if and only if $f$ is anti-monotonic and submodular.
**Hint:** Use Supplement 12 of Chapter 8.
9. Let $f : L \longrightarrow \mathbb{R}_{\geq 0}$ be a real-valued, nonnegative function, where $\mathfrak{L} = (L, \{\wedge, \vee\})$ is a lattice. Define the mapping $d : L^2 \longrightarrow \mathbb{R}_{\geq 0}$ as $d(x,y) = f(x) + f(y) - 2f(x \vee y)$ for $x,y \in L$. Prove that $d$ is a semimetric on $L$ if and only if $f$ is anti-monotonic and supramodular.

10. Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set over a set of items $I$. Prove that the mapping $d : (\mathcal{P}(I))^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by $d(H, K) = \mathit{suppcount}(H) + \mathit{suppcount}(K) - 2\mathit{suppcount}_T(HK)$ for $K, H \subseteq I$ is a semimetric on the collection of item sets.
11. Let $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ be a metric on a set $S$ and let $k$ be a number $k \in \mathbb{R}_{\geq 0}$. Prove that the function $e : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ defined by $\epsilon(x, y) = \min\{d(x, y), k\}$ is a metric on $S$.
12. Let $S$ be a set and $e : \mathcal{P}(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ be the function defined by $e(X, Y) = |X - Y|$ for $X, Y \in \mathcal{P}(S)$. Prove that $e$ satisfies the triangular axiom but fails to be a dissimilarity.
13. Let $S$ be a set and $e : S^2 \longrightarrow \mathbb{R}$ be a function such that
    - $e(x, y) = 0$ if and only if $x = y$ for $x, y \in S$,
    - $e(x, y) = e(y, x)$ for $x, y \in S$, and
    - $e(x, y) \leq e(x, z) + e(z, y)$

    for $x, y, z \in S$. Prove that $e(x, y) \geq 0$ for $x, y \in S$.
14. Let $S$ be a set and $f : S^2 \longrightarrow \mathbb{R}$ be a function such that
    - $f(x, y) = 0$ if and only if $x = y$ for $x, y \in S$;
    - $f(x, y) = f(y, x)$ for $x, y \in S$;
    - $f(x, y) \geq f(x, z) + f(z, y)$ for $x, y, z \in S$.

    Note that the triangular inequality was replaced with its inverse. Prove that the set $S$ contains at most one element.
15. Let $d : S^2 \longrightarrow \mathbb{R}$ be a function such that $d(x, y) = 0$ if and only if $x = y$ and $d(x, y) \leq d(x, z) + d(y, z)$ (note the modification of the triangular axiom), for $x, y, z \in S$. Prove that $d$ is a metric, that is, prove that $d(x, y) \geq 0$ and $d(x, y) = d(y, x)$ for all $x, y \in S$.
16. Let $f : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ be a function that satisfies the following conditions:
    a) $f(x) = 0$ if and only if $x = 0$.
    b) $f$ is monotonic on $\mathbb{R}_{\geq 0}$; that is, $x \leq y$ implies $f(x) \leq f(y)$ for $x, y \in \mathbb{R}_{\geq 0}$.
    c) $f$ is subadditive on $\mathbb{R}_{\geq 0}$; that is, $f(x+y) \leq f(x)+f(y)$ for $x, y \in \mathbb{R}_{\geq 0}$.

    a) Prove that if $d$ is a metric on a set $S$, then $fd$ is also a metric on $S$.
    b) Prove that if $d$ is a metric on $S$, the $\sqrt{d}$ and $\frac{d}{1+d}$ are also metrics on $S$. What can be said about $d^2$?
17. Let $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ be a metric on the set $S$. Prove that the function $e : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ defined by $e(x, y) = \lceil d(x, y) \rceil$ for $x, y \in S$ is also be a metric on the set $S$. Also, prove that if the ceiling function is replaced by the floor function, then this statement is no longer valid. Note that $e$ is a discretized version of the metric $d$.
18. Let $S$ be a set and let $c : S^2 \longrightarrow [0, 1]$ be a function such that $c(x, y) + c(y, x) = 1$ and $c(x, y) \leq c(x, t) + c(t, y)$ for every $x, y, t \in S$.
    a) Prove that the relation $\rho_c = \{(x, y) \in S^2 \mid c(x, y) = 1\}$ is a strict partial order on $S$.
    b) Let "$<$" be a strict partial order on a set $S$. Define the function $e : S^2 \longrightarrow \{0, \frac{1}{2}\}$ by

$$e(x,y) = \begin{cases} 1 & \text{if } x < y, \\ \frac{1}{2} & \text{if } x, y \text{ are incomparable} \\ 0 & \text{if } y < x, \end{cases}$$

for $x, y \in S$. Prove that $e(x,y) + e(y,x) = 1$ and $c(x,y) \leq c(x,t) + c(t,y)$ for every $x, y, t \in S$.

19. A function $F : \mathbb{R} \geq 0 \longrightarrow \mathbb{R}$ is *convex* if for every $s, t \in \mathbb{R}_{\geq 0}$ and $a \in [0,1]$ we have $F(as + (1-a)t) \leq aF(s) + (1-1)F(t)$.
    a) Prove that if $F(0) = 0$ and $F$ is monotonic and convex, then $F$ is subadditive.
    b) Prove that if $f$ is a metric on the set $S$, then the function given by

$$d'(x,y) = 1 - e^{-kd(x,y)},$$

where $k$ is a positive constant and $x, y \in S$, is also a metric on $S$. This metric is known as the *Schoenberg transform of* $d$ (see [37]).

**Solution:** By applying the convexity of $F$ to the interval $[0, x+y]$ with $a = \frac{x}{x+y}$, we have

$$F(a \cdot 0 + (1-a)(x+y)) \leq aF(0) + (1-a)F(x+y),$$

we have $F(y) \leq \frac{y}{x+y}F(x+y)$. Similarly, we can show that $F(x) \leq \frac{x}{x+y}F(x+y)$. By adding the last two inequalities, we obtain the subadditivity of $F$.

20. Let $S$ be a finite set and let $d : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ be a dissimilarity. Prove that there exists $a \in \mathbb{R}_{\geq 0}$ such that the dissimilarity $d_a$ defined by

$$d_a(x,y) = \begin{cases} (d(x,y))^a & \text{if } x \neq y \\ 0 & \text{if } x = y, \end{cases}$$

for $x, y \in S$ satisfies the triangular inequality.

**Hint:** Observe that $\lim_{a \to 0} d_a(x,y)$ is a dissimilarity that satisfies the triangular inequality.

21. Let $U$ be a set and let $f : U \longrightarrow S$ be an injective function. Show that if $(S,d)$ is a metric space, then the pair $(U,d')$ is also a metric space, where $d'(u,v) = d(f(u), f(v))$.
22. Let $(S_1, d_1), \ldots, (S_n, d_n)$ be $n$ metric spaces, where $n \geq 1$, and let $\nu$ be a norm on $\mathbb{R}^n$. Define the mapping $D_\nu : (S_1 \times \cdots \times S_n)^2 \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ as $D_\nu(\mathbf{x}, \mathbf{y}) = \nu(d_1(x_1, y_1), \ldots, d_n(x_n, y_n))$ for $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$.
    a) Prove that $D_\nu$ is a metric on $S_1 \times \cdots \times S_n$.
    We refer to $(S_1 \times \cdots \times S_n, D_\nu)$ as the *$\nu$-product of the metric spaces* $(S_1, d_1), \ldots, (S_n, d_n)$. When $\nu$ is the Euclidean norm, we refer to $(S_1 \times \cdots \times S_n, D_\nu)$ simply as the *product of the metric spaces* $(S_1, d_1), \ldots, (S_n, d_n)$.

b) Let $(S_1, d_1), (S_2, d_2)$ be two metric spaces. Consider the functions $\delta, \delta' : S_1 \times S_2 \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ given by

$$\delta((x, y), (u, v)) = d(x, u) + d(y, v),$$
$$\delta'((x, y), (u, v)) = \max\{d(x, u), d(y, v)\},$$

for every $(x, y), (u, v) \in S_1 \times S_2$. Prove that both $\delta$ and $\delta'$ are metrics on the set product $S_1 \times S_2$.

23. Let $(S, d)$ be a finite metric space. Prove that there exists a graph $\mathcal{G} = (S, E)$ such that $d$ is the distance associated with this graph if and only if $d(x, y) \in \mathbb{N}$ and $d(x, y) \geq 2$ implies the existence of $z \in S$ such that $d(x, y) = d(x, z) + d(z, y)$ for $x, y \in S$.
24. Prove that every metric defined on a finite set $S$ such that $|S| = 3$ is a tree metric.
25. Let $S$ be a finite set, $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ be a dissimilarity on $S$, and $s$ be an element of $S$. Define the mapping $d_{s,k} : S \times S \longrightarrow \mathbb{R}$ by

$$d_{s,k}(x, y) = \begin{cases} \frac{k + d(x,y) - d(x,s) - d(y,s)}{2} & \text{if } x \neq y \\ 0 & \text{if } x = y, \end{cases}$$

for $x, y \in S$.
    a) Prove that there is $k > 0$ such that $d_{s,k} \geq 0$ for every $s \in S$.
    b) Prove that $d$ is a tree metric if and only if there exists $k$ such that $d_{s,k}$ is an ultrametric for all $s \in S$.
26. Let $S$ be a set, $\pi$ be a partition of $S$, and $a, b$ be two numbers such that $a < b$. Prove that the mapping $d : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ given by

$$d(x, y) = \begin{cases} 0 & \text{if } x = y, \\ a & \text{if } x \neq y \text{ and } x \equiv_\pi y, \\ b & \text{if } x \not\equiv_\pi y, \end{cases}$$

is an ultrametric on $S$.
27. Prove the following extension of the statement from Exercise 26.
Let $S$ be a set, $\pi_0 < \pi_1 < \cdots < \pi_{k-1}$ be a chain of partitions on $S$, and let $a_0 < a_1 \ldots < a_{k-1} < a_k$ be a chain of positive reals. Prove that the mapping $d : S^2 \longrightarrow \mathbb{R}_{\geq 0}$ given by

$$d(x, y) = \begin{cases} 0 & \text{if } x = y, \\ a_0 & \text{if } x \neq y \text{ and } x \equiv_{\pi_0} y, \\ \vdots & \vdots \\ a_{k-1} & \text{if } x \not\equiv_{\pi_{k-2}} y \text{ and } x \equiv_{\pi_{k-1}} y, \\ a_k & \text{if } x \not\equiv_{\pi_{k-1}} y, \end{cases}$$

is an ultrametric on $S$.

**Solution:** It is clear that $d(x, y) = 0$ if and only if $x = y$ and that $d(x, y) = d(y, x)$ for any $x, y \in S$. Thus, we need to show only that $d$ satisfies the ultrametric property.

Suppose that $x, y, z \in S$ are such that $d(x, y) = a_i$ and $d(y, z) = a_j$, where $a_i < a_j$ and $i < j$. The definition of $d$ implies that $x \not\equiv_{\pi_{i-1}} y$, $x \equiv_{\pi_i} y$, and $y \not\equiv_{\pi_{j-1}} z$, $y \equiv_{\pi_j} z$. Since $\pi_i < \pi_j$, it follows that $x \equiv_{\pi_j} z$ by the transitivity of $\equiv_{\pi_j}$. Thus, $d(x, z) \leq j = \max\{d(x, y), d(y, z)\}$.

28. Using the Steinhaus transform (Lemma 10.62 on page 388), prove that if the mapping $d : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_{\geq 0}$ is defined by

$$d(\mathbf{x}, \mathbf{y}) = \frac{d_2(\mathbf{x}, \mathbf{y})}{d_2(\mathbf{x}, \mathbf{y}) + \| \mathbf{x} \| + \| \mathbf{y} \|},$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then $d$ is a metric on $\mathbb{R}^n$ such that $d(\mathbf{x}, \mathbf{y}) < 1$.

29. Using Exercises 19 and 28, prove that the mapping $e : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_{\geq 0}$ given by

$$e(\mathbf{x}, \mathbf{y}) = \frac{d_2(\mathbf{x}, \mathbf{y})}{\| \mathbf{x} \| + \| \mathbf{y} \|}$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is a metric on $\mathbb{R}^n$.

30. Prove that the following statements that concern a subset $U$ of $(\mathbb{R}^n, d_p)$ are equivalent:
    a) $U$ is bounded.
    b) There exists $n$ closed intervals $[a_1, b_1], \ldots, [a_n, b_n]$ such that $U \subseteq [a_1, b_1] \times \cdots \times [a_n, b_n]$.
    c) There exists a number $k \geq 0$ such that $d_p(\mathbf{0}, \mathbf{x}) \leq k$ for every $\mathbf{x} \in U$.

31. Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a finite subset of the metric space $(\mathbb{R}^2, d_p)$. Prove that there are at most $m$ pairs of points $(\mathbf{x}_i, \mathbf{x}_j) \in S \times S$ such that $d_p(x_i, x_j) = \mathit{diam}(S)$.

32. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$. Prove that $\mathbf{z}$ is outside the circle that has the diameter $\overline{\mathbf{x}, \mathbf{y}}$ if and only if $d_2^2(\mathbf{x}, \mathbf{z}) + d_2^2(\mathbf{y}, \mathbf{z}) > d_2^2(\mathbf{x}, \mathbf{y})$.

33. Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a finite subset of the metric space $(\mathbb{R}^2, d_p)$. The *Gabriel graph* of $S$ is the graph $\mathcal{G} = (S, E)$, where $(\mathbf{x}_i, \mathbf{x}_j) \in E$ if and only if $d_2^2(\mathbf{x}_i, \mathbf{x}_k) + d_2^2(\mathbf{x}_j, \mathbf{x}_k) > d_2^2(\mathbf{x}, \mathbf{y})$ for every $k \in \{1, \ldots, n\} - \{i, j\}$. Prove that if $\mathbf{x}, \mathbf{y}, \mathbf{z} \in S$ and $(\mathbf{y} - \mathbf{x}) \cdot (\mathbf{z} - \mathbf{x}) < 0$, for some $\mathbf{y} \in S$, then there is no edge $(\mathbf{x}, \mathbf{z})$ in the Gabriel graph of $S$. Formulate an algorithm to compute the Gabriel graph of $S$ that requires an amount of time that grows as $m^2$.

34. Let $\mathbf{C} \in \mathbb{R}^{n \times n}$ be a square matrix such that $\mathbf{Cw} = \mathbf{0}$ implies $\mathbf{w} = \mathbf{0}$ for $\mathbf{w} \in \mathbb{R}^{n \times 1}$. Define $d_{\mathbf{C}} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ by $d_{\mathbf{C}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^{tran} \mathbf{C}^{tran} \mathbf{C} (\mathbf{x} - \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n \times 1}$. Prove that $d_{\mathbf{C}}$ is a metric on $\mathbb{R}^n$.

35. Let $U_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 = 1\}$ be the set of unit vectors in $\mathbb{R}^n$. Prove that the mapping $d : U_n^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by

$$d(\mathbf{x}, \mathbf{y}) = \arccos\left(\sum_{i=1}^n x_i y_i\right),$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ belong to $U_n$, is a metric on $U_n$.

36. Let $(S, d)$ be a finite metric space. Prove that the functions $D, E : \mathcal{P}(S)^2 \longrightarrow \mathbb{R}$ defined by

$$D(U, V) = \max\{d(u, v) \mid u \in U, v \in V\},$$
$$E(U, V) = \frac{1}{|U| \cdot |V|} \sum \{d(u, v) \mid u \in U, v \in V\},$$

for $U, V \in \mathcal{P}(S)$ such that $U \neq V$, and $D(U, U) = E(U, U) = 0$ for every $U \in \mathcal{P}(S)$ are metrics on $\mathcal{P}(S)$.

37. Prove that if we replace max by min in Exercise 36, then the resulting function $F : \mathcal{P}(S)^2 \longrightarrow \mathbb{R}$ defined by

$$D(U, V) = \min\{d(u, v) \mid u \in U, v \in V\}$$

for $U, V \in \mathcal{P}(S)$ is not a metric on $\mathcal{P}(S)$, in general.

**Solution:** Let $S = U \cup V \cup W$, where

$$U = \{(0, 0), (0, 1), (1, 0), (1, 1, )\},$$
$$V = \{(2, 0), (2, 1), (2 + \ell, 0), (2 + \ell, 1)\},$$
$$W = \{(\ell + 1, 0), (\ell + 1, 1), (\ell + 2, 0), (\ell + 2, 1)\}.$$

The metric $d$ is the usual Euclidean metric in $\mathbb{R}^2$. Note that $F(U, V) = F(V, W) = 1$; however, $F(U, W) = \ell + 2$. Thus, if $\ell > 0$, the triangular axiom is violated by $F$.

38. Let $(S, d)$ be a metric space. Prove that:
    a) $d(x, T) \leq d(x, y) + d(y, T)$ for every $x, y \in S$ and $T \in \mathcal{P}(S)$.
    b) If $U$ and $V$ are nonempty subsets of $S$, then:

$$\inf_{x \in U} d(x, V) = \inf_{x \in V} d(x, U).$$

39. Let $S$ be a finite set and let $\delta : \mathcal{P}(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by

$$\delta(X, Y) = \frac{|X \oplus Y|}{|X| + |Y|}$$

for $X, Y \in \mathcal{P}(S)$. Prove that $\delta$ is a dissimilarity but not a metric.

**Hint:** Consider the set $S = \{x, y\}$ and its subsets $X = \{x\}$ and $Y = \{y\}$. Compare $\delta(X, Y)$ with $\delta(X, S) + \delta(S, Y)$.

40. Let $S$ be a finite set and let $\pi$ and $\sigma \in PART(S)$. Prove that

$$d_\beta(\pi, \sigma) \leq d_\beta(\alpha_S, \omega_S) = \mathcal{H}_\beta(\alpha_S)$$

for every $\beta \geq 1$.

41. Let $S$ be a finite set and let $\pi, \sigma$ be two partitions on $S$. Prove that if $\sigma$ covers the partition $\pi$, then there exist $B_i, B_j \in \pi$ such that

$$d_2(\pi, \sigma) = \frac{4 \cdot |B_i| \cdot |B_j|}{|S|^2}.$$

42. Let $X$ be a set of attributes of a table $\theta = (T, H, \mathbf{r})$, $\mathbf{r} = (t_1, \ldots, t_n)$, and let $\pi^X$ be the partition of $\{1, \ldots, n\}$ defined on page 295. For $\beta \in \mathbb{R}$ such that $\beta > 1$, prove that:
    a) We have $\mathcal{H}_\beta(\pi^{UV}) = \mathcal{H}_\beta(\pi^U | \pi^V) + \mathcal{H}(\pi^V)$.
    b) If $D_\beta : \mathcal{P}(H)^2 \longrightarrow \mathbb{R}_{\geq 0}$ is the semimetric defined by $D_\beta(U, V) = d_\beta(\pi^U, \pi^V)$, show that $D_\beta(U, V) = 2\mathcal{H}_\beta(\pi^{UV}) - \mathcal{H}_\beta(\pi^U) - \mathcal{H}_\beta(\pi^V)$.
    c) Prove that if $\theta$ satisfies the functional dependency $U \longrightarrow V$, then $D_\beta(U, V) = \mathcal{H}_\beta(\pi^U) - \mathcal{H}_\beta(\pi^V)$.
    d) Prove that $D_\beta(U, V) \leq \mathcal{H}_\beta(\pi^{UV}) - \mathcal{H}_\beta(\pi^U \vee \pi^V)$.
43. An attribute $A$ of a table $\theta$ is said to be *binary* if $\mathrm{Dom}(A) = \{0, 1\}$. Define the contingency matrix of two binary attributes of a table $\theta = (T, H, \mathbf{r})$ as the $2 \times 2$-matrix

$$K_{AB} = \begin{pmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{pmatrix},$$

    where $\mathbf{r} = (t_1, \ldots, t_n)$ and $n_{ij} = |\{k \mid t_k[AB] = (i, j)\}|$. Prove that $D_2(U, V) = \frac{4}{n}(n_{00} + n_{11})(n_{10} + n_{01})$, where $D_2$ is a special case of the semimetric $D_\beta$ introduced in Exercise 42.
44. Let $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$ be two partitions of a set $S$. The *Goodman-Kruskal* coefficient of $\pi$ and $\sigma$ is the number

$$GK(\pi, \sigma) = 1 - \frac{1}{|S|} \sum_{i=1}^{m} \max_{1 \leq j \leq n} |B_i \cap C_j|.$$

    a) Prove that $GK(\pi, \sigma) = 0$ if and only if $\pi \leq \sigma$.
    b) Prove that the function $GK$ is monotonic in the first argument and dually monotonic in the second argument.
    c) If $\theta, \pi, \sigma \in PART(S)$, then prove that:

$$GK(\pi \wedge \theta, \sigma) + GK(\theta, \pi) \geq GK(\theta, \pi \wedge \sigma).$$

    d) Prove that $GK(\theta, \pi) + GK(\pi, \sigma) \geq GK(\theta, \sigma)$ for $\theta, \pi, \sigma \in PART(S)$.
    e) Prove that the mapping $d_{GK} : PART(S) \times PART(S) \longrightarrow \mathbb{R}$ given by

$$d_{GK}(\pi, \sigma) = GK(\pi, \sigma) + GK(\sigma, \pi)$$

    for $\pi, \sigma \in PART(S)$, is a metric on $PART(S)$.

Partition metrics can be used to determine the validity of classification algorithms that yield essentially partitions of objects (see [72]). If $\sigma$ is a partition of a set of objects $S$ that reflects a categorization of the objects that is independent of the classification algorithm applied and $\pi$ is the partition that

is produced by the classification algorithm, then the value of $d_\beta(\pi, \sigma)$ can be used to determine the validity of $\pi$ compared to the classification of reference $\sigma$.

45. The *Rand index of two partitions* $\pi, \sigma \in PART(S)$ is the number

$$rand(\pi, \sigma) = \frac{agr(\pi, \sigma)}{\binom{|S|}{2}},$$

where $agr(\pi, \sigma)$ and $dagr(\pi, \sigma)$ were defined on page 73 . Prove that:
a) $dagr(\pi, \sigma) = \frac{|S|^2}{4} \cdot d_2(\pi, \sigma)$.
b) $rand(\pi, \sigma) = 1 - \frac{|S|^2}{2(|S|^2 - |S|)} d_2(\pi, \sigma)$.
c) $0 \leq rand(\pi, \sigma) \leq 1$; moreover, $rand(\pi, \sigma) = 1$ if and only if $\pi = \sigma$.

46. Let $S$ be a set and let $\phi : \mathbf{Seq}(S) \longrightarrow \mathbb{R}_{>0}$ be a function such that $\mathbf{u} \leq_{pref} \mathbf{v}$ implies $\phi(\mathbf{u}) \geq \phi(\mathbf{v})$ for $\mathbf{u}, \mathbf{v} \in \mathbf{Seq}(S)$.
a) Define the mapping $d_\phi : (\mathbf{Seq}(S))^2 \longrightarrow \mathbb{R}_{\geq 0}$ by

$$d_\phi(\mathbf{u}, \mathbf{v}) = \begin{cases} 0 & \text{if } \mathbf{u} = \mathbf{v}, \\ \phi(lcp(\mathbf{u}, \mathbf{v})) & \text{otherwise,} \end{cases}$$

for $\mathbf{u}, \mathbf{v} \in \mathbf{Seq}(S)$. Prove that $d_\phi$ is an ultrametric on $\mathbf{Seq}(S)$.
b) Consider an extension of the function $d_\phi$ to the set $\mathbf{Seq}_\infty(S)$ obtained by replacing the sequences $\mathbf{u}, \mathbf{v}$ in the definition of $d_\phi$ by infinite sequences. Note that this extension is possible because the longest common prefix of two distinct infinite sequences is always a finite sequence. Prove that the extended function is an ultrametric on $\mathbf{Seq}_\infty(S)$.
c) Give examples of functions $\phi$ that satisfy the conditions of Part (a) and the associated ultrametrics.

**Solution for Part (a):** It is clear that $d_\phi(\mathbf{u}, \mathbf{v}) = 0$ if and only if $\mathbf{u} = \mathbf{v}$ and that $d_\phi(\mathbf{u}, \mathbf{v}) = d_\phi(\mathbf{v}, \mathbf{u})$. Thus, we need to prove only the ultrametric inequality. Let $\mathbf{u}, \mathbf{v}, \mathbf{w}$ be three sequences. In Theorem 5.8 we have shown that at most two of the sequences $lcp(\mathbf{u}, \mathbf{v})$, $lcp(\mathbf{v}, \mathbf{w})$, $lcp(\mathbf{w}, \mathbf{u})$ are distinct and that the common value of two of these sequences is a prefix of the third sequence. This is equivalent to the ultrametric inequality for $d_\phi$.

47. Let $\mathbf{x}, \mathbf{y} \in \mathbf{Seq}_\infty(\{0, 1\})$ be two infinite binary sequences. Define $d : (\mathbf{Seq}_\infty(\{0, 1\}))^2 \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ as $d(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{\infty} \frac{|x_i - y_i|}{a^i}$, where $a > 1$.
Prove that $d$ is a metric on $\mathbf{Seq}_{infty}(\{0, 1\})$ such that $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{x}, \mathbf{z})$ implies $\mathbf{y} = \mathbf{z}$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{Seq}_{infty}(\{0, 1\})$.

A longest common subsequence of two sequences $\mathbf{x}$ and $\mathbf{y}$ is a sequence $\mathbf{z}$ that is a subsequence of both $\mathbf{x}$ and $\mathbf{y}$ and is of maximal length. For example, if $\mathbf{x} = (a_1, a_2, a_3, a_4, a_2, a_2)$ and $\mathbf{y} = (a_3, a_2, a_1, a_3, a_2, a_1, a_1, a_2, a_1)$, then both $(a_2, a_3, a_2, a_2)$ and $(a_1, a_3, a_2, a_2)$ are both longest subsequences of $\mathbf{x}$ and $\mathbf{y}$.

The length of a longest common subsequence of $\mathbf{x}$ and $\mathbf{y}$ will be denoted by $llcs(\mathbf{x}, \mathbf{y})$.

48. Let $\mathbf{x} = (x_0, \ldots, x_{n-1})$ and $\mathbf{y} = (y_0, \ldots, y_{m-1})$ be two sequences. Prove that we have

$$llcs(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{if } \mathbf{x} = \lambda \text{ or } \mathbf{y} = \lambda, \\ llcs(\mathbf{x}_{0,n-2}, \mathbf{y}_{0,m-2}) + 1 & \text{if } x_{n-1} = y_{m-1}, \\ \max\{llcs(\mathbf{x}_{0,n-2}, \mathbf{y}), llcs(\mathbf{x}, \mathbf{y}_{0,m-2})\}. \end{cases}$$

Based on this equality, formulate a tabular algorithm (similar to the one used to compute Levenshtein's distance) that can be used to compute $llcs(\mathbf{x}, \mathbf{y})$ and all longest common subsequences of $\mathbf{x}$ and $\mathbf{y}$.

49. Let $d$ be the string distance calculated with the cost scheme $(1, 1, \infty)$. Prove that $d(\mathbf{x}, \mathbf{y}) = |\mathbf{x}| + |\mathbf{y}| - 2llcs(\mathbf{x}, \mathbf{y})$.
50. Let $d$ be the string distance calculated with the cost scheme $(\infty, \infty, 1)$. Show that if $\mathbf{x} = (x_0, \ldots, x_{n-1})$ and $\mathbf{y} = (y_0, \ldots, y_{m-1})$, then

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} \infty & \text{if } |\mathbf{x}| \neq |\mathbf{y}|, \\ |\{i \mid 0 \leq i \leq n-1, x_i \neq y_i\}| & \text{if } |\mathbf{x}| = |\mathbf{y}|. \end{cases}$$

51. A *shortest common supersequence* of two sequences $\mathbf{x}$ and $\mathbf{y}$ is a sequence of minimum length that contains both $\mathbf{x}$ and $\mathbf{y}$ as subsequences. Prove that the length of a shortest common supersequence of $\mathbf{x}$ and $\mathbf{y}$ equals $|\mathbf{x}| + |\mathbf{y}| - llcs(\mathbf{x}, \mathbf{y})$.
52. Let $(S, d)$ be a finite metric space. A *metric tree for* $(S, d)$ (see [28]) is a binary tree $\mathcal{T}(S, d)$, defined as follows:
    - If $|S| = 1$, then $\mathcal{T}(S, d)$ consists of a single node that is sphere $B(s, 0)$, where $S = \{s\}$.
    - If $|S| > 1$ create a node $v$ labeled by a sphere $B(s, r) \subseteq S$ and construct the trees $\mathcal{T}(B(s, r), d)$ and $\mathcal{T}(S - B(s, r), d)$. Then, make $\mathcal{T}(B(s, r)$ and $\mathcal{T}(S - B(s, r)$ the direct descendants of $v$.

    Design an algorithm for retrieving the $k$-nearest members of $S$ to a query $q \in S$ using an existing metric tree.

The AESA algorithm (an acronym of Approximating and Eliminating Search Algorithm) starts with a finite metric space $(S, d)$, a subset $X = \{x_1, \ldots, x_n\}$ of $S$, and a query $q \in S$ and produces the nearest neighbors of $q$ in $X$. The values of distances between the members of $X$ are precomputed. The algorithm uses the semimetric $D_T$ defined in Exercise 4.

The algorithm partitions the set $X$ into three sets: $K, A, E$, where $K$ is the set of *known elements* (that is, the set of elements of $S$ for which $d(x, q)$ has been computed), $A$ is the set of *active elements*, and $E$ is the set of *eliminated elements* defined by

$$E = \{x \in X \mid D_K(x, q) > \min\{d(y, q) \mid y \in K\}\}.$$

The algorithm is given next.

**Algorithm 10.91 (AESA)**
**Input:** a metric space $(S, d)$ and a query $q \in S$;
**Output:** the set of nearest neighbors of $q$ in $X$;
**Method:** compute the matrix of dissimilarities $(d(x_i, x_j))$
of the elements of $X$;
(preprocessing phase)
$A = X$; $K = \emptyset$; $E = \emptyset$;
**while** $(A \neq \emptyset)$ **do**
$D_K(x, q) = \infty$;
select $x \in A$ such that $x = \arg\min\{D_K(x, q) \mid x \in A\}$;
compute $d(x, q)$; $K = K \cup \{x\}$; $A = A - \{x\}$;
update $r = \min\{d(x, q) \mid x \in K\}$;
update $D_K(x', q)$ for all $x' \in A$ as
$D_{K\cup\{x\}}(x', q) = \max\{D_K(x', q), |d(x, q) - d(x, x')|\}$;
$K = K \cup \{x\}$;
if $D_K(x', q) > r$ then {
$A = A - \{x'\}$;
$E = E \cup \{x'\}$;
}
**end while**;
return the set $K$

53. Prove that the AESA algorithm is indeed computing the set of nearest neighbors of $q$.
54. Let $(S, d)$ be a metric space and let $X = \{x_1, \ldots, x_n\}$ be a finite subset of $S$. Suppose that not all distances between the elements of $X$ are known. The *distance graph* of $X$ is a weighted graph $(\mathcal{G}_X, w)$, where $\mathcal{G}_X = (X, E)$. An edge $(x, y)$ exists in the underlying graph $\mathcal{G}_X$ if $d(x, y)$ is known; in this case, $w(x, y) = d(x, y)$.
If $\mathbf{p}$ is a simple path in the graph $(\mathcal{G}_X, w)$ that joins $x$ to $y$, define $\eta(\mathbf{p}) = w(\hat{e}) - \sum\{w(e) \mid e \text{ is in } \mathbf{p}, e \neq \hat{e}\}$, where $\hat{e}$ is the edge of maximum weight in $\mathbf{p}$. Prove that $d(x, y) \geq \eta(\mathbf{p})$.
55. Let *paths*$(x, y)$ be the set of simple paths in $(\mathcal{G}_X, w)$ that joins $x$ to $y$. Define the *approximate distance map* for $X$ as an $n \times n$ matrix $\mathbf{A} = (a_{ij})$ such that

$$a_{ij} = \max\{\eta(\mathbf{p}) \mid \mathbf{p} \in \textit{paths}(x_i, x_j)\}$$

for $1 \leq i, j \leq n$. Also, define the $n \times n$-matrix $\mathbf{M} = (m_{ij})$ as

$$m_{ij} = \min\{w(\mathbf{p}) \mid \mathbf{p} \in \textit{paths}(x_i, x_j)\}$$

for $1 \leq i, j \leq n$. Prove that $a_{ij} \leq d(x_i, x_j) \leq m_{ij}$ for $1 \leq i, j \leq n$.

56. Let $paths_k(x, y)$ be the set of simple paths in $(\mathcal{G}_X, w)$ that join $x$ to $y$ and do not pass through any of the vertices numbered higher than $k$, where $k \geq 1$. Denote by $a_{ij}^k, m_{ij}^k, b_{ij}^k$ the numbers

$$\begin{aligned} a_{ij}^k &= \max\{\eta(\mathbf{p}) \mid \mathbf{p} \in paths_k(x_i, x_j)\}, \\ m_{ij}^k &= \min\{w(\mathbf{p}) \mid \mathbf{p} \in paths_k(x_i, x_j)\}, \\ b_{ij}^k &= \max\{\eta(\mathbf{p}) \mid \mathbf{p} \in paths_k(x_i, x_k) paths_k(x_k, x_j)\} \end{aligned}$$

for $1 \leq i, j \leq n$. Prove that

$$b_{ij}^k = \max\{a_{ij}^{k-1} - m_{ij}^{k-1}, a_{jk}^{k-1} - m_{ki}^{k-1}\}$$

for $1 \leq k \leq n$.

## Bibliographical Comments

A comprehensive, research-oriented source for metric spaces is [37].

The reader should consult two excellent surveys [27, 17] on searches in metric spaces. Nearest-neighbor searching is comprehensively examined in [28],[29]. The AESA algorithm was introduced in [140, 141]. Weighted graphs of partially defined metric spaces and approximate distance maps are defined and studied in [143, 122], where Exercises 54–56 originate. Finally, we refer the reader to two fundamental references for all things about metrics [37],[36].

# 11

# Topologies and Measures on Metric Spaces

## 11.1 Introduction

The study of topological properties of metric spaces allows us to present an introduction to the dimension theory of these spaces, a topic that is relevant for data mining due to its role in understanding the complexity of searching in data sets that have a natural metric structure.

Topologies of metric spaces are presented in Section 11.2.

## 11.2 Metric Space Topologies

Metrics spaces are naturally equipped with topologies using a mechanism that we describe next.

**Theorem 11.1.** *Let $(S, d)$ be a metric space. The family of sets $\mathcal{O}_d$ defined by*

$$\mathcal{O}_d = \{L \in \mathcal{P}(S) \mid \text{ for each } x \in L \text{ there exists } \epsilon > 0 \text{ such that } C_d(x, \epsilon) \subseteq L\}$$

*is a topology on the set $S$.*

*Proof.* We have $\emptyset \in \mathcal{O}_d$ because there is no $x$ in $\emptyset$, so the condition of the definition of $\mathcal{O}_d$ is vacuously satisfied. The set $S$ belongs to $\mathcal{O}_d$ because $C_d(x, \epsilon) \subseteq S$ for every $x \in S$ and every positive number $\epsilon$.

If $\{U_i \mid i \in I\} \subseteq \mathcal{O}_d$ and $x \in \bigcup\{U_i \mid i \in I\}$, then $x \in U_j$ for some $j \in I$. Then, there exists $\epsilon > 0$ such that $S(x, \epsilon) \subseteq U_j$ and therefore $S(x, \epsilon) \subseteq \bigcup\{U_i \mid i \in I\}$. Thus, $\bigcup\{U_i \mid i \in I\} \in \mathcal{O}_d$.

Finally, let $U, V \in \mathcal{O}_d$ and let $x \in U \cap V$. Since $U \in \mathcal{O}_d$, there exists $\epsilon > 0$ such that $C(x, \epsilon) \subseteq U$. Similarly, there exists $\epsilon'$ such that $C(x, \epsilon') \subseteq V$. If $\epsilon_1 = \min\{\epsilon, \epsilon'\}$, then

$$C(x, \epsilon_1) \subseteq C(x, \epsilon) \cap C(x, \epsilon') \subseteq U \cap V,$$

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_11, © Springer-Verlag London Limited 2008

so $U \cap V \in \mathcal{O}_d$. This concludes the argument. □

Theorem 11.1 justifies the following definition.

**Definition 11.2.** *Let $d$ be a metric on a set $S$. The* topology induced by $d$ *is the family of sets*

$$\mathcal{O}_d = \{L \in \mathcal{P}(S) \mid \text{for each } x \in L \text{ there exists } \epsilon > 0 \text{ such that } C_d(x, \epsilon) \subseteq L\}.$$

*We refer to the pair $(S, \mathcal{O}_d)$ as a* topological metric space.

*Example 11.3.* The usual topology of the set of real numbers $\mathbb{R}$ introduced in Example 6.4 is actually induced by the metric $d : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}_{\geq 0}$ given by $d(x, y) = |x - y|$ for $x, y \in \mathbb{R}$. Recall that, by Theorem 6.10, every open set of this space is the union of a countable set of disjoint open intervals.

The next statement explains the terms "open sphere" and "closed sphere," which we have used previously.

**Theorem 11.4.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. If $t \in S$ and $r > 0$, then any open sphere $C(t, r)$ is an open set and any closed sphere $B(t, r)$ is a closed set in the topological space $(S, \mathcal{O}_d)$.*

*Proof.* Let $x \in C(t, r)$, so $d(t, x) < r$. Choose $\epsilon$ such that $\epsilon < r - d(t, x)$. We claim that $C(x, \epsilon) \subseteq C(t, r)$. Indeed, let $z \in C(x, \epsilon)$. We have $d(x, z) < \epsilon < r - d(t, x)$. Therefore, $d(z, t) \leq d(z, x) + d(x, t) < r$, so $z \in C(t, r)$, which implies $C(x, \epsilon) \subseteq C(t, r)$. We conclude that $C(t, r)$ is an open set.

To show that the closed sphere $B(t, r)$ is a closed set, we will prove that its complement $S - B(t, r) = \{u \in S \mid d(u, t) > r\}$ is an open set. Let $v \in S - B(t, r)$. Now choose $\epsilon$ such that $\epsilon < d(v, t) - r$. It is easy to see that $C(v, \epsilon) \subseteq S - B(t, r)$, which proves that $S - B(t, r)$ is an open set. □

**Corollary 11.5.** *The collection of all open spheres in a topological metric space $(S, \mathcal{O}_d)$ is a basis.*

*Proof.* This statement follows immediately from Theorem 11.4. □

The definition of open sets in a topological metric space implies that a subset $L$ of a topological metric space $(S, \mathcal{O}_d)$ is closed if and only if for every $x \in S$ such that $x \notin L$ there is $\epsilon > 0$ such that $C(x, \epsilon)$ is disjoint from $L$. Thus, if $C(x, \epsilon) \cap L \neq \emptyset$ for every $\epsilon > 0$ and $L$ is a closed set, then $x \in L$.

The closure and the interior operators $\mathbf{K}_{\mathcal{O}_d}$ and $\mathbf{I}_{\mathcal{O}_d}$ in a topological metric space $(S, \mathcal{O}_d)$ are described next.

**Theorem 11.6.** *In a topological metric space $(S, \mathcal{O}_d)$, we have*

$$\boldsymbol{K}_{\mathcal{O}_d}(U) = \{x \in S \mid C(x, \epsilon) \cap U \neq \emptyset \text{ for every } \epsilon > 0\}$$

*and*

$$\boldsymbol{I}_{\mathcal{O}_d}(U) = \{x \in S \mid C(x, \epsilon) \subseteq U \text{ for some } \epsilon > 0\}$$

*for every $U \in \mathcal{P}(S)$.*

*Proof.* Let $\mathbf{K} = \mathbf{K}_{\mathcal{O}_d}$. If $C(x,\epsilon) \cap U \neq \emptyset$ for every $\epsilon > 0$, then clearly $C(x,\epsilon) \cap \mathbf{K}(U) \neq \emptyset$ for every $\epsilon > 0$ and therefore $x \in \mathbf{K}(U)$ by a previous observation.

Now let $x \in \mathbf{K}(U)$ and let $\epsilon > 0$. Suppose that $C(x,\epsilon) \cap U = \emptyset$. Then, $U \subseteq S - C(x,\epsilon)$ and $S - C(x,\epsilon)$ is a closed set. Therefore, $\mathbf{K}(U) \subseteq S - C(x,\epsilon)$. This is a contradiction because $x \in \mathbf{K}(U)$ and $x \notin S - C(x,\epsilon)$.

The second part of the theorem follows from the first part and from Corollary 6.27. □

If the metric topology $\mathcal{O}_d$ is clear from the context, then we will denote the closure operator $\mathbf{K}_{\mathcal{O}_d}$ simply by $\mathbf{K}$.

**Corollary 11.7.** *The subset $U$ of the topological metric space $(S, \mathcal{O}_d)$ is closed if and only if $C(x,\epsilon) \cap U \neq \emptyset$ for every $\epsilon > 0$ implies $x \in U$.*

*Proof.* This statement is an immediate consequence of Theorem 11.6. □

**Corollary 11.8.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $L$ be a subset of $K$. Then, the border $\partial L$ of the set $L$ is given by*

$$\partial L = \{x \in S | \text{ for every } \epsilon > 0, C(x,\epsilon) \cap L \neq \emptyset, \text{ and } C(x,\epsilon) \cap (S - L) \neq \emptyset\}.$$

**Theorem 11.9.** *Let $T$ be a subset of a topological metric space $(S, \mathcal{O}_d)$. We have $diam(T) = diam(\boldsymbol{K}(T))$.*

*Proof.* Since $T \subseteq \mathbf{K}(T)$, it follows immediately that $diam(T) \leq diam(\mathbf{K}(T))$, so we have to prove only the reverse inequality.

Let $u, v \in \mathbf{K}(T)$. For every positive number $\epsilon$, we have $C(u,\epsilon) \cap T \neq \emptyset$ and $C(v,\epsilon) \cap T \neq \emptyset$. Thus, there exists $x, y \in T$ such that $d(u,x) < \epsilon$ and $d(v,y) < \epsilon$. Thus, $d(u,v) \leq d(u,x) + d(x,y) + d(y,v) \leq 2\epsilon + diam(T)$ for every $\epsilon$, which implies $d(u,v) \leq diam(T)$ for every $u, v \in \mathbf{K}(T)$. This yields $diam(\mathbf{K}(T)) \leq diam(T)$. □

A metric topology can be defined, as we shall see, by more than one metric.

**Definition 11.10.** *Two metrics $d$ and $d'$ defined on a set $S$ are* topologically equivalent *if the topologies $\mathcal{O}_d$ and $\mathcal{O}_{d'}$ are equal.*

*Example 11.11.* Let $d$ and $d'$ be two metrics defined on a set $S$. If there exist two numbers $a, b \in \mathbb{R}_{>0}$ such that

$$ad(x,y) \leq d'(x,y) \leq bd(x,y),$$

for $x, y \in S$, then $\mathcal{O}_d = \mathcal{O}'_d$. Let $C_d(x,r)$ be an open sphere centered in $x$, defined by $d$. The inequalities above imply

$$C_d\left(\frac{r}{b}\right) \subseteq C_{d'}(x,r) \subseteq C_d\left(x, \frac{r}{a}\right).$$

Let $L \in \mathcal{O}_d$. By Definition 11.2, for each $x \in L$ there exists $\epsilon > 0$ such that $C_d(x,\epsilon) \subseteq L$. Then, $C'_d(x, a\epsilon) \subseteq C_d(x,\epsilon) \subseteq L$, which implies $L \in \mathcal{O}_{d'}$. We leave it to the reader to prove the reverse inclusion $\mathcal{O}_{d'} \subseteq \mathcal{O}_d$.

By Corollary 10.52, any two Minkowski metrics $d_p$ and $d_q$ on $\mathbb{R}^n$ are topologically equivalent.

## 11.3 Continuous Functions in Metric Spaces

Continuous functions between topological spaces were introduced in Definition 6.62.

Next, we give a characterization of continuous functions between topological metric spaces.

**Theorem 11.12.** *Let $(S, \mathcal{O}_d)$ and $(T, \mathcal{O}_e)$ be two topological metric spaces. The following statements concerning a function $f : S \longrightarrow T$ are equivalent:*
(i) *$f$ is a continuous function.*
(ii) *For every $x \in S$ and $\epsilon > 0$, there exists $\delta > 0$ such that*

$$f(C_d(x, \delta)) \subseteq C_e(f(x), \epsilon).$$

*Proof.* (i) implies (ii): Suppose that $f$ is a continuous function. Since $C_e(f(x), \epsilon)$ is an open set in $(T, \mathcal{O}_e)$, the set $f^{-1}(C_e(f(x), \epsilon)$ is an open set in $(S, \mathcal{O}_d)$. Clearly, $x \in f^{-1}(C_e(f(x), \epsilon))$, so by the definition of the metric topology there exists $\delta > 0$ such that $C_d(x, \delta) \subseteq f^{-1}(C_e(f(x), \epsilon)$, which yields $f(C_d(x, \delta)) \subseteq C_e(f(x), \epsilon)$.

(ii) implies (i): Let $V$ be an open set of $(T, \mathcal{O}_e)$. If $f^{-1}(V)$ is empty, then it is clearly open. Therefore, we may assume that $f^{-1}(V)$ is not empty. Let $x \in f^{1-}(V)$. Since $f(x) \in V$ and $V$ is open, there exists $\epsilon > 0$ such that $C_e(f(x), \epsilon) \subseteq V$. By Part (ii) of the theorem, there exists $\delta > 0$ such that $f(C_d(x, \delta)) \subseteq C_e(f(x), \epsilon)$, which implies $x \in C_d(x, \delta) \subseteq f^{-1}(V)$. This means that $f^{-1}(V)$ is open, so $f$ is continuous. □

In general, for a continuous function $f : S \longrightarrow T$, the number $\delta$ depends both on $x$ and on $\epsilon$. If $\delta$ is dependent only on $\epsilon$, then we say that $f$ is *uniformly continuous*. Thus, $f$ is uniformly continuous if for every $\epsilon > 0$ there exists $\delta$ such that if $d(u, v) < \delta$, then $e(f(u), f(v)) < \epsilon$.

**Theorem 11.13.** *Let $(S, \mathcal{O}_d)$ and $(T, \mathcal{O}_e)$ be two topological metric spaces and let $f : S \longrightarrow T$ be a function. The following statements are equivalent.*
(i) *$f$ is uniformly continuous.*
(ii) *For all sequences $\boldsymbol{u} = (u_0, u_1, \ldots)$ and $\boldsymbol{v} = (v_0, v_1, \ldots)$ in $\mathbf{Seq}_\infty(S)$ such that $\lim_{n\to\infty} d(u_n, v_n) = 0$ we have $\lim_{n\to\infty} e(f(u_n), f(v_n)) = 0$.*
(iii) *For all sequences $\boldsymbol{u} = (u_0, u_1, \ldots)$ and $\boldsymbol{v} = (v_0, v_1, \ldots)$ in $\mathbf{Seq}_\infty(S)$ such that $\lim_{n\to\infty} d(u_n, v_n) = 0$, we have $\lim_{k\to\infty} e(f(u_{n_k}), f(v_{n_k})) = 0$, where $(u_{n_0}, u_{n_1}, \ldots)$ and $(v_{n_0}, v_{n_1}, \ldots)$ are two arbitrary subsequences of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively.*

*Proof.* (i) implies (ii): For $\epsilon > 0$, there exists $\delta$ such that $d(u, v) < \delta$ implies $e(f(u), f(v)) < \epsilon$. Therefore, if $\mathbf{u}$ and $\mathbf{v}$ are sequences as above, there exists $n_\delta$ such that $n > n_\delta$ implies $d(u_n, v_n) < \delta$, so $e(f(u_n), f(v_n)) < \epsilon$. Thus, $\lim_{n\to\infty} e(f(u_n), f(v_n)) = 0$.

(ii) implies (iii): This implication is obvious.

(iii) implies (i): Suppose that $f$ satisfies (iii) but is not uniformly continuous. Then, there exists $\epsilon > 0$ such that for every $\delta > 0$ there exist $u, v \in X$ such that $d(u, v) < \delta$ and $e(f(u), f(v)) > \epsilon$. Let $u_n, v_n$ be such that $d(u_n, v_n) < \frac{1}{n}$ for $n \geq 1$. Then, $\lim_{n\to\infty} d(u_n, v_n) = 0$ but $e(f(u_n), f(v_n))$ does not converge to 0. □

*Example 11.14.* The function $f : \mathbb{R} \longrightarrow \mathbb{R}$ given by $f(x) = x \sin x$ is continuous but not uniformly continuous. Indeed, let $u_n = n\pi$ and $v_n = n\pi + \frac{1}{n}$. Note that $\lim_{n\to\infty} |u_n - v_n| = 0$, $f(u_n) = 0$, and $f(v_n) = (n\pi + \frac{1}{n}) \sin(n\pi + \frac{1}{n}) = (n\pi + \frac{1}{n})(-1)^n \sin \frac{1}{n}$. Therefore,

$$\begin{aligned} &\lim_{n\to\infty} |f(u_n) - f(v_n)| \\ &= \lim_{n\to\infty} \left(n\pi + \frac{1}{n}\right) \sin \frac{1}{n} \\ &= \pi \lim_{n\to\infty} \frac{n}{\sin \frac{1}{n}} = \pi. \end{aligned}$$

This implies that $f$ is not uniformly continuous.

A local continuity property is introduced next.

**Definition 11.15.** *Let $(S, \mathcal{O}_d)$ and $(T, \mathcal{O}_e)$ be two topological metric spaces and let $x \in S$.*

*A function $f : S \longrightarrow T$ is* continuous in $x$ *if for every $\epsilon > 0$ there exists $\delta > 0$ such that*

$$f(C(x, \delta)) \subseteq C(f(x), \epsilon).$$

It is clear that $f$ is continuous if it is continuous in every $x \in S$.

The definition can be restated by saying that $f$ is continuous in $x$ if for every $\epsilon > 0$ there is $\delta > 0$ such that $d(x, y) < \delta$ implies $e(f(x), f(y)) < \epsilon$.

## 11.4 Separation Properties of Metric Spaces

We begin by discussing properties of distances between points and sets in metric spaces.

**Theorem 11.16.** *Let $(S, d)$ be a metric space. The following statements hold:*
(i) $|d(u, V) - d(u', V)| \leq d(u, u')$,
(ii) $d(u, V) = 0$ *if and only if* $u \in \boldsymbol{K}(V)$, *and*
(iii) $d(u, V) = d(u, \boldsymbol{K}(V))$
*for every $u, u' \in S$ and $V \subseteq S$.*

*Proof.* Let $v$ be an element of $v$. Since $d(u, v) \leq d(u, u') + d(u', v)$, it follows that $d(u, V) \leq d(u, u') + d(u', v)$ by the definition of $d(u, V)$. Since $d(u, V) - d(u, u') \leq d(u', v)$ for every $v \in V$, by the same definition we obtain $d(u, V) -$

$d(u, u') \leq d(u', V)$. Thus, $d(u, V) - d(u', V) \leq d(u, u')$. Reversing the roles of $u$ and $u'$ we have $d(u', V) - d(u, V) \leq d(u, u')$, which gives Inequality (i).

Suppose that $d(u, V) = 0$. Again, by the definition of $d(u, V)$, for every $\epsilon > 0$ there exists $v \in V$ such that $d(u, v) < \epsilon$, which means that $C(u, \epsilon) \cap V \neq \emptyset$. By Theorem 11.6, we have $u \in \mathbf{K}(V)$. The converse implication is immediate, so (ii) holds.

Finally, to prove (iii), observe that $V \subseteq \mathbf{K}(V)$ implies that $d(u, \mathbf{K}(V)) \leq d(u, V)$, so we need to prove only the reverse inequality.

Let $w$ be an arbitrary element of $\mathbf{K}(V)$. By Theorem 11.6, for every $\epsilon > 0$, $C(w, \epsilon) \cap V \neq \emptyset$. Let $v \in C(w, \epsilon) \cap V$. We have

$$d(u, v) \leq d(u, w) + d(w, v) \leq d(u, w) + \epsilon,$$

so $d(u, V) \leq d(u, w) + \epsilon$. Since this inequality holds for every $\epsilon$, $d(u, V) \leq d(u, w)$ for every $w \in \mathbf{K}(V)$, so $d(u, V) \leq d(u, \mathbf{K}(V))$. This allows us to conclude that $d(u, V) = d(u, \mathbf{K}(V))$. □

Theorem 11.16 can be restated using the function $d_V : S \longrightarrow \mathbb{R}_{\geq 0}$ defined by $d_V(u) = d(u, V)$ for $u \in S$. Thus, for every subset $V$ of $S$, we have

$$\begin{aligned} &|d_V(u) - d_V(u')| \leq d(u, u'), \\ &d_V(u) = 0 \text{ if and only if } u \in \mathbf{K}(V), \text{ and} \\ &d_V = d_{\mathbf{K}(V)} \end{aligned}$$

for $u, u' \in S$.

**Corollary 11.17.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $V$ be a subset of $S$. Then, the function $d_V$ is a continuous function on the space $(S, \mathcal{O}_d)$ and $\boldsymbol{K}(V) = d_V^{-1}(0)$.*

*Proof.* This statement is an immediate consequence of Theorem 11.16. □

The notions of an open sphere and a closed sphere in a metric space $(S, d)$ are extended by defining the sets $C(T, r)$ and $B(T, r)$ as

$$\begin{aligned} C(T, r) &= \{u \in S \mid d(u, T) < r\}, \\ B(T, r) &= \{u \in S \mid d(u, T) \leq r\}, \end{aligned}$$

for $T \in \mathcal{P}(S)$ and $r \geq 0$, respectively.

The next statement is a generalization of Theorem 11.4.

**Theorem 11.18.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. For every set $T$, $T \subseteq S$, and every $r > 0$, $C(T, r)$ is an open set and $B(T, r)$ is a closed set in $(S, \mathcal{O}_d)$.*

*Proof.* Let $u \in C(T, r)$. We have $d(u, T) < r$, or, equivalently, $\inf\{d(u, t) \mid t \in T\} < r$. We claim that if $\epsilon$ is a positive number such that $\epsilon < \frac{r}{2}$, then $C(u, \epsilon) \subseteq C(T, r)$.

Let $z \in C(u, \epsilon)$. For every $v \in T$, we have $d(z, v) \leq d(z, u) + d(u, v) < \epsilon + d(u, v)$. From the definition of $d(u, T)$ as an infimum, it follows that there exists $v' \in T$ such that $d(u, v') < d(u, V) + \frac{\epsilon}{2}$, so $d(z, v') < d(u, T) + \epsilon < r + \epsilon$. Since this inequality holds for every $\epsilon > 0$, it follows that $d(z, v') < r$, so $d(z, T) < r$, which proves that $C(u, \epsilon) \subseteq C(T, r)$. Thus, $C(T, r)$ is an open set.

Suppose now that $s \in \mathbf{K}(B(T, r))$. By Part (ii) of Theorem 11.16, we have $d(s, B(T, r)) = 0$, so $\inf\{d(s, w) \mid w \in B(T, r)\} = 0$. Therefore, for every $\epsilon > 0$, there is $w \in B(T, r)$ such that $d(s, w) < \epsilon$. Since $d(w, T) \leq r$, it follows from the first part of Theorem 11.16 that $|d(s, T) - d(w, T)| \leq d(s, w) < \epsilon$ for every $\epsilon > 0$. This implies $d(s, T) = d(w, T)$, so $s \in B(T, r)$. This allows us to conclude that $B(T, r)$ is indeed a closed set. □

**Theorem 11.19 (Lebesgue's Lemma).** *Let $(S, \mathcal{O}_d)$ be a topological metric space that is compact and let $\mathcal{C}$ be an open cover of this space. There exists $r \in \mathbb{R}_{>0}$ such that for every subset $U$ with $diam(U) < r$ there is a set $L \in \mathcal{C}$ such that $U \subseteq L$.*

*Proof.* Suppose that the statement is not true. Then, for every $k \in \mathbb{P}$, there exists a subset $U_k$ of $S$ such that $diam(U_k) < \frac{1}{k}$ and $U_k$ is not included in any of the sets $L$ of $\mathcal{C}$. Since $(S, \mathcal{O}_d)$ is compact, there exists a finite subcover $\{L_1, \ldots, L_p\}$ of $\mathcal{C}$.

Let $x_{ik}$ be an element in $U_k - L_i$. For every two points $x_{ik}, x_{jk}$, we have $d(x_{ik}, x_{jk}) \leq \frac{1}{k}$ because both belong to the same set $U_k$. By Theorem 11.60, the compactness of $S$ implies that any sequence $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots)$ contains a convergent subsequence. Denote by $x_i$ the limit of this subsequence, where $1 \leq i \leq p$. The inequality $d(x_{ik}, x_{jk}) \leq \frac{1}{k}$ for $k \geq 1$ implies that $d(x_i, x_j) = 0$ so $x_i = x_j$ for $1 \leq i, j \leq p$. Let $x$ be their common value. Then $x$ does not belong to any of the sets $L_i$, which contradicts the fact that $\{L_1, \ldots, L_p\}$ is an open cover. □

**Theorem 11.20.** *Every topological metric space $(S, \mathcal{O}_d)$ is a Hausdorff space.*

*Proof.* Let $x$ and $y$ be two distinct elements of $S$, so $d(x, y) > 0$. Choose $\epsilon = \frac{d(x,y)}{3}$. It is clear that for the open spheres $C(x, \epsilon)$ and $C(y, \epsilon)$, we have $x \in C(x, \epsilon)$, $y \in C(y, \epsilon)$, and $C(x, \epsilon) \cap C(y, \epsilon) = \emptyset$, so $(S, \mathcal{O}_d)$ is indeed a Hausdorff space. □

**Corollary 11.21.** *Every compact subset of a topological metric space is closed.*

*Proof.* This follows directly from Theorems 11.20 and 6.89. □

**Corollary 11.22.** *If $S$ is a finite set and $d$ is a metric on $S$, then the topology $\mathcal{O}_d$ is the discrete topology.*

*Proof.* Let $S = \{x_1, \ldots, x_n\}$ be a finite set. We saw that every singleton $\{x_i\}$ is a closed set. Therefore, every subset of $S$ is closed as a finite union of closed sets. □

**Theorem 11.23.** *Every topological metric space $(S, \mathcal{O}_d)$ is a $T_4$ space.*

*Proof.* We need to prove that for all disjoint closed sets $H_1$ and $H_2$ of $S$ there exist two open disjoint sets $V_1$ and $V_2$ such that $H_1 \subseteq V_1$ and $H_2 \subseteq V_2$.

Let $x \in H_1$. Since $H_1 \cap H_2 = \emptyset$, it follows that $x \notin H_2 = \mathbf{K}(H_2)$, so $d(x, H_2) > 0$ by Part (ii) of Theorem 11.16. By Theorem 11.18, the set $C\left(H_1, \frac{d(x,L)}{3}\right)$ is an open set and so is

$$Q_H = \bigcup \left\{ C\left(H_1, \frac{d(x,L)}{3}\right) \mid x \in H_1 \right\}.$$

The open set $Q_{H_2}$ is defined in a similar manner as

$$Q_{H_2} = \bigcup \left\{ C\left(H_2, \frac{d(y,H_1)}{3}\right) \mid y \in H_2 \right\}.$$

The sets $Q_{H_1}$ and $Q_{H_2}$ are disjoint because $t \in Q_{H_1} \cap Q_{H_2}$ implies that there is $x_1 \in H_1$ and $x_2 \in H_2$ such that $d(t, x_1) < \frac{d(x_1,H_2)}{3}$ and $d(t, x_2) < \frac{d(x_2,H_1)}{3}$. This, in turn, would imply

$$d(x_1, x_2) < \frac{d(x_1, H_2) + d(x_2, H_1)}{3} \leq \frac{2}{3} d(x_1, x_2),$$

which is a contradiction. Therefore, $(S, \mathcal{O}_d)$ is a $T_4$ topological space. □

**Corollary 11.24.** *Every metric space is normal.*

*Proof.* By Theorem 11.20, a metric space is a $T_2$ space and therefore a $T_1$ space. The statement then follows directly from Theorem 11.23. □

**Corollary 11.25.** *Let $H$ be a closed set and $L$ be an open set in a topological metric space $(S, \mathcal{O}_d)$ such that $H \subseteq L$. Then, there is an open set $V$ such that $H \subseteq V \subseteq \boldsymbol{K}(V) \subseteq L$.*

*Proof.* The closed sets $H$ and $S - L$ are disjoint. Therefore, since $(S, \mathcal{O})$ is normal, there exist two disjoint open sets $V$ and $W$ such that $H \subseteq V$ and $S - L \subseteq W$. Since $S - W$ is closed and $V \subseteq S - W$, it follows that $\mathbf{K}(V) \subseteq S - W \subseteq L$. Thus, we obtain $H \subseteq V \subseteq \mathbf{K}(V) \subseteq L$. □

A stronger form of Theorem 11.23, where the disjointness of the open sets is replaced by the disjointness of their closures, is given next.

**Theorem 11.26.** *Let $(S, \mathcal{O}_d)$ be a metric space. For all disjoint closed sets $H_1$ and $H_2$ of $S$, there exist two open sets $V_1$ and $V_2$ such that $H_1 \subseteq V_1$, $H_2 \subseteq V_2$, and $\boldsymbol{K}(V_1) \cap \boldsymbol{K}(V_2) = \emptyset$.*

*Proof.* By Theorem 11.23, we obtain the existence of the disjoint open sets $Q_{H_1}$ and $Q_{H_2}$ such that $H_1 \subseteq Q_{H_1}$ and $H_2 \subseteq Q_{H_2}$. We claim that the closures of these sets are disjoint.

Suppose that $s \in \mathbf{K}(Q_{H_1}) \cap \mathbf{K}(Q_{H_2})$. Then, we have $C\left(s, \frac{\epsilon}{12}\right) \cap Q_{H_1} \neq \emptyset$ and $C\left(s, \frac{\epsilon}{12}\right) \cap Q_{H_2} \neq \emptyset$. Thus, there exist $t \in Q_{H_1}$ and $t' \in Q_{H_2}$ such that $d(t, s) < \frac{\epsilon}{12}$ and $d(t', s) < \frac{\epsilon}{12}$.

As in the proof of the previous theorem, there is $x_1 \in H_1$ and $y_1 \in H_2$ such that $d(t, x_1) < \frac{d(x_1, H_2)}{3}$ and $d(t', y_1) < \frac{d(y_1, H_1)}{3}$. Choose $t$ and $t'$ above for $\epsilon = d(x_1, y_1)$. This leads to a contradiction because

$$d(x_1, y_1) \leq d(x_1, t) + d(t, s) + d(s, t') + d(t', y_1) \leq \frac{5}{6} d(x_1, y_1).$$

□

**Corollary 11.27.** *Let $(S, \mathcal{O}_d)$ be a metric space. If $x \in L$, where $L$ is an open subset of $S$, then there exists two open sets $V_1$ and $V_2$ in $S$ such that $x \in V_1$, $S - L \subseteq V_2$, and $\boldsymbol{K}(V_1) \cap \boldsymbol{K}(V_2) = \emptyset$.*

*Proof.* The statement follows by applying Theorem 11.26 to the disjoint closed sets $H_1 = \{x\}$ and $H_2 = S - L$. □

Recall that the Bolzano-Weierstrass property of topological spaces was introduced in Theorem 6.61. Namely, a topological space $(S, \mathcal{O})$ has the Bolzano-Weierstrass property if every infinite subset $T$ of $S$ has at least one accumulation point. For metric spaces, this property is equivalent to compactness, as we show next.

**Theorem 11.28.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. The following three statements are equivalent:*
(i) *$(S, \mathcal{O}_d)$ is compact.*
(ii) *$(S, \mathcal{O}_d)$ has the Bolzano-Weierstrass property.*
(iii) *Every countable open cover of $(S, \mathcal{O}_d)$ contains a finite subcover.*

*Proof.* (i) implies (ii): by Theorem 6.61.

(ii) implies (iii): Let $\{L_n \mid n \in \mathbb{N}\}$ be a countable open cover of $S$. Without loss of generality, we may assume that none of the sets $L_n$ is included in $\bigcup_{p=1}^{n-1} L_p$; indeed, if this is not the case, we can discard $L_n$ and still have a countable open cover. Let $x_n \in L_n - \bigcup_{p=1}^{n-1} L_p$ and let $U = \{x_n \mid n \in \mathbb{N}\}$. Since $(S, \mathcal{O}_d)$ has the Bolzano-Weierstrass property, we have $U' \neq \emptyset$, so there exists an accumulation point $z$ of $U$. In every open set $L$ that contains $z$, there exists $x_n \in U$ such that $x_n \neq z$.

Since $\{L_n \mid n \in \mathbb{N}\}$ is an open cover, there exists $L_m$ such that $z \in L_m$. Suppose that the set $L_m$ contains only a finite number of elements $x_{n_1}, \ldots, x_{n_k}$, and let $d = \min\{d(z, x_{n_i}) \mid 1 \leq i \leq k\}$. Then, $L_m \cap C\left(z, \frac{d}{2}\right)$ is an open set that contains no elements of $U$ with the possible exception of $z$, which contradicts the fact that $z$ is an accumulation point. Thus, $L_m$

contains an infinite subset of $U$, which implies that there exists $x_q \in L_m$ for some $q > m$. This contradicts the definition of the elements $x_n$ of $U$. We conclude that there exists a number $r_0$ such that $L_r - \bigcup_{i=0}^{r-1} L_i = \emptyset$ for $r \geq r_0$, so $S = L_0 \cup \cdots \cup L_{r_0-1}$, which proves that $L_0, \ldots, L_{r_0-1}$ is a finite subcover.

(iii) implies (i). Let $\epsilon$ be a positive number. Suppose that there is an infinite sequence $\mathbf{x} = (x_0, \ldots, x_n, \ldots)$ such that $d(x_i, x_j) > \epsilon$ for every $i, j \in \mathbb{N}$ such that $i \neq j$. Consider the open spheres $C(x_i, \epsilon)$ and the set

$$C = S - \mathbf{K}\left(\bigcup_{i \in \mathbb{N}} C\left(x_i, \frac{\epsilon}{2}\right)\right).$$

We will show that $\{C\} \cup \{C(x_i, \epsilon) \mid i \in \mathbb{N}\}$ is a countable open cover of $S$.

Suppose that $x \in S - C$; that is $x \in \mathbf{K}\left(\bigcup_{i \in \mathbb{N}} C\left(x_i, \frac{\epsilon}{2}\right)\right)$. By Theorem 6.37, we have either that $x \in \bigcup_{i \in \mathbb{N}} C\left(x_i, \frac{\epsilon}{2}\right)$ or $x$ is an accumulation point of that set.

In the first case, $x \in \bigcup_{i \in \mathbb{N}} C(x_i, \epsilon)$ because $C\left(x_i, \frac{\epsilon}{2}\right) \subseteq C(x_i, \epsilon)$. If $x$ is an accumulation point of $\bigcup_{i \in \mathbb{N}} C\left(x_i, \frac{\epsilon}{2}\right)$, given any open set $L$ such that $x \in L$, then $L$ must intersect at least one of the spheres $C\left(x_i, \frac{\epsilon}{2}\right)$. Suppose that $C\left(x, \frac{\epsilon}{2}\right) \cap C\left(x_i, \frac{\epsilon}{2}\right) \neq \emptyset$, and let $t$ be a point that belongs to this intersection. Then, $d(x, x_i) < d(x, t) + d(t, x_i) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$, so $x \in C(x_i, \epsilon)$.

Therefore, $\{C\} \cup \{C(x_i, \epsilon) \mid i \in \mathbb{N}\}$ is a countable open cover of $S$. Since every countable open cover of $(S, \mathcal{O}_d)$ contains a finite subcover, it follows that this open cover contains a finite subcover. Observe that there exists an open sphere $C(x_i, \epsilon)$ that contains infinitely many $x_n$ because none of these elements belongs to $C$. Consequently, for any two of these points, the distance is less than $\epsilon$, which contradicts the assumption we made initially about the sequence $\mathbf{x}$.

Choose $\epsilon = \frac{1}{k}$ for some $k \in \mathbb{N}$ such that $k \geq 1$. Since there is no infinite sequence of points such that every two distinct points are at a distance greater than $\frac{1}{k}$, it is possible to find a finite sequence of points $\mathbf{x} = (x_0, \ldots, x_{n-1})$ such that $i \neq j$ implies $d(x_i, x_j) > \frac{1}{k}$ for $0 \leq i, j \leq n-1$ and for every other point $x \in S$ there exists $x_i$ such that $d(x_i, x) \leq \frac{1}{k}$.

Define the set $L_{k,m,i}$ as the open sphere $C\left(x_i, \frac{1}{m}\right)$, where $x_i$ is one of the points that belongs to the sequence above determined by $k$ and $m \in \mathbb{N}$ and $m \geq 1$. The collection $\{L_{k,m,i} \mid m \geq 1, 0 \leq i \leq n-1\}$ is clearly countable. We will prove that each open set of $(S, \mathcal{O}_d)$ is a union of sets of the form $L_{k,m,i}$; in other words, we will show that this family of sets is a basis for $(S, \mathcal{O}_d)$.

Let $L$ be an open set and let $z \in L$. Since $L$ is open, there exists $\epsilon > 0$ such that $z \in C(z, \epsilon) \subseteq L$. Choose $k$ and $m$ such that $\frac{1}{k} < \frac{1}{m} < \frac{\epsilon}{2}$. By the definition of the sequence $\mathbf{x}$, there is $x_i$ such that $d(z, x_i) < \frac{1}{k}$. We claim that

$$L_{k,m,i} = C\left(x_i, \frac{1}{m}\right) \subseteq L.$$

Let $y \in L_{k,m,i}$. Since $d(z,y) \leq d(z,x_i) + d(x_i,y) < \frac{1}{k} + \frac{1}{m} < \epsilon$, it follows that $L_{k,m,i} \subseteq C(z,\epsilon) \subseteq L$. Since $d(y,z) < \frac{1}{k} < \frac{1}{m}$, we have $z \in L_{k,m,i}$. This shows that $L$ is a union of sets of the form $L_{k,m,i}$, so this family of sets is a countable open cover of $S$. It follows that that there exists a finite open cover of $(S, \mathcal{O}_d)$ because every countable open cover of $(S, \mathcal{O}_d)$ contains a finite subcover. □

Closed or open spheres in ultrametric spaces have an interesting property, which we discuss next.

**Corollary 11.29.** *If $d$ is an ultrametric on $S$, then any closed sphere $B(t,r)$ and any open sphere $C(t,r)$ is a clopen set in the topological ultrametric space $(S, \mathcal{O}_d)$.*

*Proof.* We already know that $B(t,r)$ is closed. To prove that this set is also open if $d$ is an ultrametric, let $s \in B(t,r)$. We saw that $s$ is again a center of the sphere (see Theorem 10.22). Therefore, $C\left(s, \frac{r}{2}\right) \subseteq B(t,r)$, so $B(t,r)$ is open. We leave the proof that $C(t,r)$ is also closed to the reader. □

By Theorem 6.34, the border of a closed sphere or of an open sphere in an ultrametric space is empty.

**Theorem 11.30.** *Let $d$ and $d'$ be two metrics on a set $S$ such that there exist $c_0, c_1 \in \mathbb{R}_{>0}$ for which $c_0 d(x,y) \leq d'(x,y) \leq c_1 d(x,y)$ for every $x, y \in S$. Then, the topologies $\mathcal{O}_d$ and $\mathcal{O}_{d'}$ coincide.*

*Proof.* Suppose that $L \in \mathcal{O}_d$, and let $x \in L$. There exists $\epsilon > 0$ such that $C_d(x,\epsilon) \subseteq L$. Note that $C_{d'}(x, c_1\epsilon) \subseteq C_d(x,\epsilon)$. Thus, $C_{d'}(x,\epsilon') \subseteq L$, where $\epsilon' = c_1\epsilon$, which shows that $L \in \mathcal{O}_{d'}$. In a similar manner, one can prove that $\mathcal{O}_{d'} \subseteq \mathcal{O}_d$, so the two topologies are equal. □

If $d$ and $d'$ are two metrics on a set $S$ such that $\mathcal{O}_d = \mathcal{O}_{d'}$, we say that $d$ and $d'$ are topologically equivalent. Corollary 10.55 implies that all metrics $d_p$ on $\mathbb{R}^n$ with $p \geq 1$ are topologically equivalent.

In Section 6.2, we saw that if a topological space has a countable basis, then the space is separable (Theorem 6.46) and each open cover of the basis contains a countable subcover (Corollary 6.49). For metric spaces, these properties are equivalent, as we show next.

**Theorem 11.31.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. The following statements are equivalent:*
(i) *$(S, \mathcal{O}_d)$ has a countable basis.*
(ii) *$(S, \mathcal{O}_d)$ is a separable.*
(iii) *Every open cover of $(S, \mathcal{O}_d)$ contains a countable subcover.*

*Proof.* By Theorem 6.46 and Corollary 6.49, the first statement implies (ii) and (iii). Therefore, it suffices to prove that (iii) implies (ii) and (ii) implies (i).

To show that (iii) implies (ii), suppose that every open cover of $(S, \mathcal{O}_d)$ contains a countable subcover. The collection of open spheres $\{C\left(x, \frac{1}{n}\right) \mid x \in$

$S, n \in \mathbb{N}_{>0}\}$ is an open cover of $S$ and therefore there exists a countable set $T_n \subseteq S$ such that $\mathcal{C}_n = \{C\left(x, \frac{1}{n}\right) \mid x \in T_n, n \in \mathbb{N}_{>0}\}$ is an open cover of $S$. Let $C = \bigcup_{n \geq 1} T_n$. By Theorem 1.130, $C$ is a countable set.

We claim that $C$ is dense in $(S, \mathcal{O}_d)$. Indeed, let $s \in S$ and choose $n$ such that $n > \frac{1}{\epsilon}$. Since $\mathcal{C}_n$ is an open cover of $S$, there is $x \in T_n$ such that $s \in C\left(x, \frac{1}{n}\right) \subseteq C(x, \epsilon)$. Since $T_n \subseteq C$, it follows that $C$ is dense in $(S, \mathcal{O}_d)$. Thus, $(S, \mathcal{O}_d)$ is separable.

To prove that (ii) implies (i), let $(S, \mathcal{O}_d)$ be a separable space. There exists a countable set $U$ that is dense in $(S, \mathcal{O}_d)$. Consider the countable collection

$$\mathcal{C} = \left\{ C\left(u, \frac{1}{n}\right) \mid u \in U, n \geq 1 \right\}.$$

If $L$ is an open set in $(S, \mathcal{O}_d)$ and $x \in L$, then there exists $\epsilon > 0$ such that $C(x, \epsilon) \subseteq L$. Let $n$ be such that $n > \frac{2}{\epsilon}$. Since $U$ is dense in $(S, \mathcal{O}_d)$, we know that $x \in \mathbf{K}(U)$, so there exists $y \in S(x, \epsilon) \cap U$ and $x \in C\left(y, \frac{1}{n}\right) \subseteq C\left(x, \frac{2}{n}\right) \subseteq C(x, \epsilon) \subseteq L$. Thus, $\mathcal{C}$ is a countable basis. □

**Theorem 11.32.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. Every closed set of this space is a countable intersection of open sets, and every open set is a countable union of closed sets.*

*Proof.* Let $H$ be a closed set and let $U_n$ be the open set

$$U_n = \bigcup_{n \geq 1} \left\{ C\left(x, \frac{1}{n}\right) \mid x \in F \right\}.$$

It is clear that $H \subseteq \bigcap_{n \geq 1} U_n$. Now let $u \in \bigcap_{n \geq 1} U_n$ and let $\epsilon$ be an arbitrary positive number. For every $n \geq 1$, there is an element $x_n \in H$ such that $d(u, x_n) < \frac{1}{n}$. Thus, if $\frac{1}{n} < \epsilon$, we have $x_n \in H \cap C(u, \epsilon)$, so $C(u, \epsilon) \cap H \neq \emptyset$. By Corollary 11.7, it follows that $u \in H$, which proves the reverse inclusion $\bigcap_{n \geq 1} U_n \subseteq H$. This shows that every closed set is a countable union of open sets.

If $L$ is an open set, then its complement is closed and, by the first part of the theorem, it is a countable intersection of open sets. Thus, $L$ itself is a countable union of closed sets. □

**Definition 11.33.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. A $G_\delta$-*set *is a countable intersection of open sets. An $F_\delta$-*set *is a countable union of open sets.*

Now, Theorem 11.32 can be restated by saying that every closed set of a topological metric space is a $G_\delta$-set and every open set is an $F_\delta$-set.

**Theorem 11.34.** *Let $U$ be a $G_\delta$-set in the topological metric space $(S, \mathcal{O}_d)$. If $T$ is a $G_\delta$-set in the subspace $U$, then $T$ is a $G_\delta$-set in $S$.*

*Proof.* Since $T$ is a $G_\delta$-set in the subspace $U$, we can write $T = \bigcap_{n\in\mathbb{N}} L_n$, where each $L_n$ is an open set in the subspace $U$. By the definition of the subspace topology, for each $L_n$ there exists an open set in $S$ such that $L_n = L'_n \cap U$, so

$$T = \bigcap_{n\in\mathbb{N}} L_n = \bigcap_{n\in\mathbb{N}} (L'_n \cap U) = U \cap \bigcap_{n\in\mathbb{N}} L'_n.$$

Since $U$ is a countable intersection of open sets of $S$, the last equality shows that $T$ is a countable intersection of open sets of $S$ and hence a $G_\delta$-set in $S$. □

## 11.5 Sequences in Metric Spaces

**Definition 11.35.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $\boldsymbol{x} = (x_0, \ldots, x_n, \ldots)$ be a sequence in $\mathbf{Seq}_\infty(S)$.*

*The sequence $\boldsymbol{x}$* converges *to an element $x$ of $S$ if for every $\epsilon > 0$ there exists $n_\epsilon \in \mathbb{N}$ such that $n \geq n_\epsilon$ implies $x_n \in C(x, \epsilon)$.*

*A sequence $\boldsymbol{x}$ is* convergent *if it converges to an element $x$ of $S$.*

**Theorem 11.36.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $\boldsymbol{x} = (x_0, \ldots, x_n, \ldots)$ be a sequence in $\mathbf{Seq}_\infty(S)$. If $\boldsymbol{x}$ is convergent, then there exists a unique $x$ such that $\boldsymbol{x}$ converges to $x$.*

*Proof.* Suppose that there are two distinct elements $x$ and $y$ of the set $S$ that satisfy the condition of Definition 11.35. We have $d(x, y) > 0$. Define $\epsilon = \frac{d(x,y)}{3}$. By definition, there exists $n_\epsilon$ such that $n \geq n_\epsilon$ implies $d(x, x_n) < \epsilon$ and $d(x_n, y) < \epsilon$. By applying the triangular inequality, we obtain

$$d(x, y) \leq d(x, x_n) + d(x_n, y) < 2\epsilon = \frac{2}{3} d(x, y),$$

which is a contradiction. □

If the sequence $\mathbf{x} = (x_0, \ldots, x_n, \ldots)$ converges to $x$, this is denoted by $\lim_{n\to\infty} x_n = x$.

An alternative characterization of continuity of functions can be formulated using convergent sequences.

**Theorem 11.37.** *Let $(S, \mathcal{O}_d)$ and $(T, \mathcal{O}_e)$ be two topological metric spaces and let $f : S \longrightarrow T$. The function $f$ is continuous in $x$ if and only if for every sequence $\boldsymbol{x} = (x_0, \ldots, x_n, \ldots)$ such that $\lim_{n\to\infty} x_n = x$ we have $\lim_{n\to\infty} f(x_n) = f(x)$.*

*Proof.* Suppose that $f$ is continuous in $x$, and let $\mathbf{x} = (x_0, \ldots, x_n, \ldots)$ be a sequence such that $\lim_{n\to\infty} x_n = x$. Let $\epsilon > 0$. By Definition 11.35, there exists $\delta > 0$ such that $f(C(x, \delta)) \subseteq C(f(x), \epsilon)$. Since $\lim_{n\to\infty} x_n = x$, there exists $n_\delta$

such that $n \geq n_\delta$ implies $x_n \in C(x, \delta)$. Then, $f(x_n) \in f(C(x, \delta)) \subseteq C(f(x), \epsilon)$. This shows that $\lim_{n\to\infty} f(x_n) = f(x)$.

Conversely, suppose that for every sequence $\mathbf{x} = (x_0, \ldots, x_n, \ldots)$ such that $\lim_{n\to\infty} x_n = x$, we have $\lim_{n\to\infty} f(x_n) = f(x)$. If $f$ were not continuous in $x$, we would have an $\epsilon > 0$ such that for all $\delta > 0$ we would have $y \in C(x, \delta)$ but $f(y) \notin C(f(x), \epsilon)$. Choosing $\delta = \frac{1}{n}$, let $y_n \in S$ such that $y_n \in C\left(x, \frac{1}{n}\right)$ and $f(y_n) \notin C(f(x), \epsilon)$. This yields a contradiction because we should have $\lim_{n\to\infty} f(y_n) = f(x)$. □

### Sequences of Real Numbers

**Theorem 11.38.** *Let $\boldsymbol{x} = (x_0, \ldots, x_n, \ldots)$ be a sequence in $(\mathbb{R}, \mathcal{O})$, where $\mathcal{O}$ is the usual topology on $\mathbb{R}$.*

*If $\boldsymbol{x}$ is an increasing (decreasing) sequence and there exists a number $b \in \mathbb{R}$ such that $x_n \leq b$ ($x_n \geq b$, respectively), then the sequence $\boldsymbol{x}$ is convergent.*

*Proof.* Since the set $\{x_n \mid n \in \mathbb{N}\}$ is bounded above, its supremum $s$ exists by the Completeness Axiom for $\mathbb{R}$ given in Section 4.4. We claim that $\lim_{n\to\infty} x_n = s$. Indeed, by Theorem 4.28, for every $\epsilon > 0$ there exists $n_\epsilon \in \mathbb{N}$ such that $s - \epsilon < x_{n_\epsilon} \leq s$. Therefore, by the monotonicity of the sequence and its boundedness, we have $s - \epsilon < x_n \leq s$ for $n \geq n_\epsilon$, so $x_n \in C(x, \epsilon)$, which proves that $\mathbf{x}$ converges to $s$.

We leave it to the reader to show that any decreasing sequence in $(\mathbb{R}, \mathcal{O})$ that is bounded below is convergent. □

If $\mathbf{x}$ is an increasing sequence and there is no upper bound for $\mathbf{x}$, this means that for every $b \in \mathbb{R}$ there exists a number $n_b$ such that $n \geq n_b$ implies $x_n > b$. If this is the case, we say that $\mathbf{x}$ is a sequence *divergent to* $+\infty$ and we write $\lim_{n\to\infty} x_n = +\infty$. Similarly, if $\mathbf{x}$ is a decreasing sequence and there is no lower bound for it, this means that for every $b \in \mathbb{R}$ there exists a number $n_b$ such that $n \geq n_b$ implies $x_n < b$. In this case, we say that $\mathbf{x}$ is a sequence *divergent to* $-\infty$ and we write $\lim_{n\to\infty} x_n = -\infty$.

Theorem 11.38 and the notion of a divergent sequence allow us to say that $\lim_{n\to\infty} x_n$ exists for every increasing or decreasing sequence; this limit may be a real number or $\pm\infty$ depending on the boundedness of the sequence.

**Theorem 11.39.** *Let $[a_0, b_0] \supseteq [a_1, b_1] \supset \cdots \supset [a_n, b_n] \supset \cdots$ be a sequence of nested closed intervals of real numbers. There exists a closed interval $[a, b]$ such that $a = \lim_{n\to\infty} a_n$, $b = \lim_{n\to\infty} b_n$, and*

$$[a, b] = \bigcap_{n\in\mathbb{N}} [a_n, b_n].$$

*Proof.* The sequence $a_0, a_1, \ldots, a_n, \ldots$ is clearly increasing and bounded because we have $a_n \leq b_m$ for every $n, m \in \mathbb{N}$. Therefore, it converges to a number $a \in \mathbb{R}$ and $a \leq b_m$ for every $m \in \mathbb{N}$. Similarly, $b_0, b_1, \ldots, b_n, \ldots$ is

a decreasing sequence that is bounded below, so it converges to a number $b$ such that $a_n \leq b$ for $n \in \mathbb{N}$. Consequently, $[a, b] \subseteq \bigcap_{n\in\mathbb{N}}[a_n, b_n]$.

Conversely, let $c$ be a number in $\bigcap_{n\in\mathbb{N}}[a_n, b_n]$. Since $c \geq a_n$ for $n \in \mathbb{N}$, it follows that $c \geq \sup\{a_n \mid n \in \mathbb{N}\}$, so $c \geq a$. A similar argument shows that $c \leq b$, so $c \in [a, b]$, which implies the reverse inclusion $\bigcap_{n\in\mathbb{N}}[a_n, b_n] \subseteq [a, b]$. □

In Example 6.56, we saw that every closed interval $[a, b]$ of $\mathbb{R}$ is a compact set. This allows us to prove the next statement.

**Theorem 11.40 (Bolzano-Weierstrass Theorem).** *A bounded sequence of real numbers has a convergent subsequence.*

*Proof.* Let $\mathbf{x} = (x_0, \ldots, x_n, \ldots)$ be a bounded sequence of real numbers. The boundedness of $\mathbf{x}$ implies the existence of a closed interval $D_0 = [a_0, b_0]$ such that $\{x_n \mid n \in \mathbb{N}\} \subseteq [a_0, b_0]$.

Let $c = \frac{a_0+b_0}{2}$ be the midpoint of $D_0$. At least one of the sets $\mathbf{x}^{-1}([a_0, c_0])$, $\mathbf{x}^{-1}([c_0, b_0])$ is infinite. Let $[a_1, b_1]$ be one of $[a_0, c_0]$ or $[c_0, b_0]$, for which $\mathbf{x}^{-1}([a_0, c_0])$, $\mathbf{x}^{-1}([c_0, b_0])$ is infinite.

Suppose that we have constructed the interval $D_n = [a_n, b_n]$ having $c_n = \frac{a_n+b_n}{2}$ as its midpoint such that $\mathbf{x}^{-1}(D_n)$ is infinite. Then, $D_{n+1} = [a_{n+1}, b_{n+1}]$ is obtained from $D_n$ as one of the intervals $[a_n, c_n]$ or $[c_n, b_n]$ that contains $x_n$ for infinitely many $n$.

Thus, we obtain a descending sequence of closed intervals $[a_0, b_0] \supset [a_1, b_1] \supset \cdots$ such that each interval contains an infinite set of members of the sequence $\mathbf{x}$. By Theorem 11.39, we have $[a, b] = \bigcup_{n\in\mathbb{N}}[a_n, b_n]$, where $a = \lim_{n\to\infty} a_n$ and $b = \lim_{n\to\infty} b_n$. Note that $b_n - a_n = \frac{b_0-a_0}{2^n}$, so $a = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = b$.

The interval $D_0$ contains at least one member of $\mathbf{x}$, say $x_{n_0}$. Since $D_1$ contains infinitely many members of $\mathbf{x}$, there exists a member $x_{n_1}$ of $\mathbf{x}$ such that $n_1 > n_0$. Continuing in this manner, we obtain a subsequence $x_{n_0}, x_{n_1}, \ldots, x_{n_p}, \ldots$. Since $a_p \leq x_{n_p} \leq b_p$, it follows that the sequence $(x_{n_0}, x_{n_1}, \ldots, x_{n_p}, \ldots)$ converges to $a$. □

Let $\mathbf{x} = (x_0, x_1, \ldots)$ be a sequence of real numbers. Consider the sequence of sets $S_n = \{x_n, x_{n+1}, \ldots\}$ for $n \in \mathbb{N}$. It is clear that $S_0 \supseteq S_1 \supseteq \cdots \subseteq S_n \supseteq \cdots$. Therefore, we have the increasing sequence of numbers $\inf S_0 \leq \inf S_1 \leq \cdots \leq \inf S_n \leq \cdots$; we define $\liminf \mathbf{x}$ as $\lim_{n\to\infty} \inf S_n$. On the other hand, we have the decreasing sequence $\sup S_0 \geq \sup S_1 \geq \cdots \geq \sup S_n \geq \cdots$ of numbers; we define $\limsup \mathbf{x}$ as $\lim_{n\to\infty} \sup S_n$.

*Example 11.41.* Let $\mathbf{x}$ be the sequence defined by $x_n = (-1)^n$ for $n \in \mathbb{N}$. It is clear that $\sup S_n = 1$ and $\inf S_n = -1$. Therefore, $\limsup \mathbf{x} = 1$ and $\liminf \mathbf{x} = -1$.

**Theorem 11.42.** *For every sequence $\boldsymbol{x}$ of real numbers, we have* $\liminf \boldsymbol{x} \leq \limsup \boldsymbol{x}$.

*Proof.* Let $S_n = \{x_n, x_{n+1}, \ldots\}$, $y_n = \inf S_n$, and $z_n = \sup S_n$ for $n \in \mathbb{N}$. If $p \geq n$, we have $y_n \leq y_p \leq z_p \leq z_n$, so $y_n \leq z_p$ for every $n, p$ such that $p \geq n$. Since $z_1 \geq z_2 \geq \cdots \geq z_p$, it follows that $y_n \leq z_p$ for every $p \in \mathbb{N}$. Therefore, $\limsup \mathbf{x} = \lim_{p\to\infty} z_p \geq y_n$ for every $n \in \mathbb{N}$, which in turn implies $\liminf \mathbf{x} = \lim_{n\to\infty} y_n \leq \limsup \mathbf{x}$. □

**Corollary 11.43.** *Let $\boldsymbol{x} = (x_0, x_1, \ldots, x_n, \ldots)$ be a sequence of real numbers. We have $\liminf \boldsymbol{x} = \limsup \boldsymbol{x} = \ell$ if and only if $\lim_{n\to\infty} x_n = \ell$.*

*Proof.* Suppose that $\liminf \mathbf{x} = \limsup \mathbf{x} = \ell$ and that it is not the case that $\lim_{n\to\infty} x_n = \ell$. This means that there exists $\epsilon > 0$ such that, for every $m \in \mathbb{N}$, $n \geq m$ implies $|x_n - \ell| \geq \epsilon$, which is equivalent to $x_n \geq \ell + \epsilon$ or $x_n \leq \ell - \epsilon$. Thus, at least one of the following cases occurs:

(i) there are infinitely many $n$ such that $x_n \geq \ell + \epsilon$, which implies that $\limsup x_n \geq \ell + \epsilon$, or

(ii) there are infinitely many $n$ such that $x_n \leq \ell - \epsilon$, which implies that $\liminf x_n \geq \ell - \epsilon$.

Either case contradicts the hypothesis, so $\lim_{n\to\infty} x_n = \ell$.

Conversely, suppose that $\lim_{n\to\infty} x_n = \ell$. There exists $n_\epsilon$ such that $n \geq n_\epsilon$ implies $\ell - \epsilon < x_n < \ell + \epsilon$. Thus, $\sup\{x_n \mid n \geq n_\epsilon\} \leq \ell + \epsilon$, so $\limsup \mathbf{x} \leq \ell + \epsilon$. Similarly, $y - \epsilon \leq \liminf \mathbf{x}$ and the inequality

$$\ell - \epsilon \leq \liminf \mathbf{x} \leq \limsup \mathbf{x} \leq \ell + \epsilon,$$

which holds for every $\epsilon > 0$, implies $\liminf \mathbf{x} = \limsup \mathbf{x} = \ell$. □

### Sequences and Open and Closed Sets

**Theorem 11.44.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. A subset $U$ of $S$ is open if and only if for every $x \in U$ and every sequence $(x_0, \ldots, x_n, \ldots)$ such that $\lim_{n\to\infty} x_n = x$ there is $m$ such that $n \geq m$ implies $x_n \in U$.*

*Proof.* Suppose $U$ is an open set. Since $x \in U$, there exists $\epsilon > 0$ such that $C(x, \epsilon) \subseteq U$. Let $(x_0, \ldots, x_n, \ldots)$ be such that $\lim_{n\to\infty} = x$. By Definition 11.35, there exists $n_\epsilon$ such that $n \geq n_\epsilon$ implies $x_n \in C(x, \epsilon) \subseteq U$.

Conversely, suppose that the condition is satisfied and that $U$ is not open. Then, there exists $x \in U$ such that for every $n \geq 1$ we have $C\left(x, \frac{1}{n}\right) - U \neq \emptyset$. Choose $x_{n-1} \in C\left(x, \frac{1}{n}\right)$ for $n \geq 1$. It is clear that the sequence $(x_0, \ldots, x_n, \ldots)$ converges to $x$. However, none of the members of this sequence belong to $U$. This contradicts our supposition, so $U$ must be an open set. □

**Theorem 11.45.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. A subset $W$ of $S$ is closed if and only if for every sequence $\boldsymbol{x} = (x_0, \ldots, x_n, \ldots) \in \mathbf{Seq}_\infty(W)$ such that $\lim_{n\to\infty} x_n = x$ we have $x \in W$.*

*Proof.* If $W$ is a closed set and $\mathbf{x} = (x_0, \ldots, x_n, \ldots)$ is a sequence whose members belong to $W$, then none of these members belong to $S - W$. Since $S - W$ is an open set, by Theorem 11.45, it follows that $x \notin S - W$; that is, $x \in W$.

Conversely, suppose that for every sequence $(x_0, \ldots, x_n, \ldots)$ such that $\lim_{n\to\infty} = x$ and $x_n \in W$ for $n \in \mathbb{N}$ we have $x \in W$. Let $v \in S - W$, and suppose that for every $n \geq 1$ the open sphere $C\left(v, \frac{1}{n}\right)$ is not included in $S - W$. This means that for each $n \geq 1$ there is $z_{n-1} \in C\left(v, \frac{1}{n}\right) \cap W$. We have $\lim_{n\to\infty} z_n = v$; this implies $v \in W$. This contradiction means that there is $n \geq 1$ such that $C\left(v, \frac{1}{n}\right) \subseteq V$, so $V$ is an open set. Consequently, $W = S - V$ is a closed set. □

## 11.6 Completeness of Metric Spaces

Let $\mathbf{x} = (x_0, \ldots, x_n, \ldots)$ be a sequence in the topological metric space $(S, \mathcal{O}_d)$ such that $\lim_{n\to\infty} x_n = x$. If $m, n > n_{\frac{\epsilon}{2}}$, we have $d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$. In other words, if $\mathbf{x}$ is a sequence that converges to $x$, then given a positive number $\epsilon$ we have members of the sequence closer than $\epsilon$ if we go far enough in the sequence. This suggests the following definition:

**Definition 11.46.** *A sequence $\boldsymbol{x} = (x_0, \ldots, x_n, \ldots)$ in the topological metric space $(S, \mathcal{O}_d)$ is a* Cauchy sequence *if for every $\epsilon > 0$ there exists $n_\epsilon \in \mathbb{N}$ such that $m, n \geq n_\epsilon$ implies $\rho(x_m, x_n) < \epsilon$.*

**Theorem 11.47.** *Every convergent sequence in a topological metric space $(S, \mathcal{O}_d)$ is a Cauchy sequence.*

*Proof.* Let $\mathbf{x} = (x_0, x_1, \ldots)$ be a convergent sequence and let $x = \lim_{n\to\infty} \mathbf{x}$. There exists $n'_{\frac{\epsilon}{2}}$ such that if $n > n'_{\frac{\epsilon}{2}}$, then $d(x_n, x) < \frac{\epsilon}{2}$. Thus, if $m, n \geq n_\epsilon = n'_{\frac{\epsilon}{2}}$, it follows that

$$d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

which means that $\mathbf{x}$ is a Cauchy sequence. □

*Example 11.48.* The converse of Theorem 11.47 is not true, in general, as we show next.

Let $((0, 1), d)$ be the metric space equipped with the metric $d(x, y) = |x - y|$ for $x, y \in (0, 1)$. The sequence defined by $x_n = \frac{1}{n+1}$ for $n \in \mathbb{N}$ is a Cauchy sequence. Indeed, it suffices to take $m, n \geq \frac{1}{\epsilon} - 1$ to obtain $|x_n - x_m| < \epsilon$; however, the sequence $x_n$ is not convergent to an element of $(0, 1)$.

**Definition 11.49.** *A topological metric space is* complete *if every Cauchy sequence is convergent.*

*Example 11.50.* The topological metric space $(\mathbb{R}, \mathcal{O}_d)$, where $d(x, y) = |x - y|$ for $x, y \in\in \mathbb{R}$, is complete.

Let $\mathbf{x} = (x_0, x_1, \ldots)$ be a Cauchy sequence in $\mathbb{R}$. For every $\epsilon > 0$, there exists $n_\epsilon \in \mathbb{N}$ such that $m, n \geq n_\epsilon$ implies $|x_m - x_n| < \epsilon$. Choose $m_0 \in \mathbb{N}$ such that $m_0 \geq n_\epsilon$. Thus, if $n \geq n_\epsilon$, then $x_{m_0} - \epsilon < x_n < x_{m_0} + \epsilon$, which means that $\mathbf{x}$ is a bounded sequence. By Theorem 11.40, the sequence $\mathbf{x}$ contains a bounded subsequence $(x_{i_0}, x_{i_1}, \ldots)$ that is convergent. Let $\ell = \lim_{k\to\infty} x_{i_k}$. It is not difficult to see that $\lim_{x_n} x_n = \ell$, which shows that $(\mathbb{R}, \mathcal{O}_d)$ is complete.

**Theorem 11.51.** *Let $(S, \mathcal{O}_d)$ be a complete topological metric space. If $T$ is a closed subset of $S$, then the subspace $T$ is complete.*

*Proof.* Let $T$ be a closed subset of $S$ and let $\mathbf{x} = (x_0, x_1, \ldots)$ be a Cauchy sequence in this subspace. The sequence $\mathbf{x}$ is a Cauchy sequence in the complete space $S$, so there exists $x = \lim_{n\to\infty} x_n$. Since $T$ is closed, we have $x \in T$, so $T$ is complete.

Conversely, suppose that $T$ is complete. Let $x \in \mathbf{K}(T)$. There exists a sequence $\mathbf{x} = (x_0, x_1, \ldots) \in \mathbf{Seq}_\infty(T)$ such that $\lim_{n\to\infty} x_n = x$. Then, $\mathbf{x}$ is a Cauchy sequence in $T$, so there is a limit $t$ of this sequence in $T$. The uniqueness of the limit implies $x = t \in T$, so $T$ is a closed set. □

**Theorem 11.52.** *There is no clopen set in the topological space $(\mathbb{R}, \mathcal{O})$ except the empty set and the set $\mathbb{R}$.*

*Proof.* Suppose that $L$ is a clopen subset of $\mathbb{R}$ that is distinct from $\emptyset$ and $\mathbb{R}$. Then, there exist $x \in L$ and $y \notin L$. Starting from $x$ and $y$, we define inductively the terms of two sequences $\mathbf{x} = (x_0, \ldots, x_n, \ldots)$ and $\mathbf{y} = (y_0, \ldots, y_n, \ldots)$ as follows. Let $x_0 = x$ and $y_0 = y$. Suppose that $x_n$ and $y_n$ are defined. Then,

$$x_{n+1} = \begin{cases} \frac{x_n + y_n}{2} & \text{if } \frac{x_n + y_n}{2} \in L, \\ x_n & \text{otherwise,} \end{cases}$$

and

$$y_{n+1} = \begin{cases} \frac{x_n + y_n}{2} & \text{if } \frac{x_n + y_n}{2} \notin L, \\ y_n & \text{otherwise.} \end{cases}$$

It is clear that $\{x_n \mid n \in \mathbb{N}\} \subseteq L$ and $\{y_n \mid n \in \mathbb{N}\} \subseteq \mathbb{R} - L$. Moreover, we have

$$|y_{n+1} - x_{n+1}| = \frac{|y_n - x_n|}{2} = \cdots = \frac{|y - x|}{2^{n+1}}.$$

Note that

$$|x_{n+1} - x_n| \leq |y_n - x_n| \leq \frac{|y - x|}{2^n}.$$

This implies that $\mathbf{x}$ is a Cauchy sequence and therefore there is $x = \lim_{n\to\infty} x_n$; moreover, the sequence $\mathbf{y}$ also converges to $x$, so $x$ belongs to $\partial L$, which is a contradiction. □

**Theorem 11.53.** *In a complete topological metric space $(S, \mathcal{O}_d)$, every descending sequence of closed sets $V_0 \supset V_1 \supset \cdots V_n \subset V_{n+1} \supset \cdots$ such that $\lim_{n\to\infty} diam(V_n) = 0$ has a nonempty intersection, that is, $\bigcap_{n\in\mathbb{N}} V_n \neq \emptyset$.*

*Proof.* Consider a sequence $x_0, x_1, \ldots, x_n, \ldots$ such that $x_n \in V_n$. This is a Cauchy sequence. Indeed, let $\epsilon > 0$. Since $\lim_{n\to\infty} diam(V_n) = 0$, there exists $n_\epsilon$ such that if $m, n > n_\epsilon$ we have $x_m, x_n \in V_{\min\{m,n\}}$. Since $\min\{m, n\} \geq n_\epsilon$, it follows that $d(x_m, x_n) \leq diam(V_{\min m,n}) < \epsilon$. Since the space $(S, \mathcal{O}_d)$ is complete, it follows that there exists $x \in S$ such that $\lim_{n\to\infty} x_n = x$. Note that all members of the sequence above belong to $V_m$, with the possible exception of the first $m$ members. Therefore, by Theorem 11.45, $x \in V_m$, so $x \in \bigcap_{n\in\mathbb{N}} V_n$, so $\bigcap_{n\in\mathbb{N}} V_n \neq \emptyset$. □

Recall that the definition of Baire spaces was introduced on page 233.

**Theorem 11.54.** *Every complete topological metric space is a Baire space.*

*Proof.* We prove that if $(S, \mathcal{O}_d)$ is complete, then it satisfies the first condition of Theorem 6.28.

Let $L_1, \ldots, L_n, \ldots$ be a sequence of open subsets of $S$ that are dense in $S$ and let $L$ be an open, nonempty subset of $S$. We construct inductively a sequence of closed sets $H_1, \ldots, H_n, \ldots$ that satisfy the following conditions:

(i) $H_1 \subseteq L_0 \cap L$,
(ii) $H_n \subseteq L_n \cap H_{n-1}$ for $n \geq 2$,
(iii) $\mathbf{I}(H_n) \neq \emptyset$, and
(iv) $diam(H_n) \leq \frac{1}{n}$

for $n \geq 2$.

Since $L_1$ is dense in $S$, by Theorem 6.22, $L_1 \cap L \neq \emptyset$, so there is a closed sphere of diameter less than 1 enclosed in $L_1 \cap L$. Define $H_1$ as this closed sphere.

Suppose that $H_{n-1}$ was constructed. Since $\mathbf{I}(H_{n-1}) \neq \emptyset$, the open set $L_n \cap \mathbf{I}(H_{n-1})$ is not empty because $L_n$ is dense in $S$. Thus, there is a closed sphere $H_n$ included in $L_n \cap \mathbf{I}(H_{n-1})$, and therefore included in $L_n \cap H_{n-1}$, such that $diam(H_n) < \frac{1}{n}$. Clearly, we have $\mathbf{I}(H_n) \neq \emptyset$. By applying Theorem 11.53 to the descending sequence of closed sets $H_1, \ldots, H_n, \ldots$, the completeness of the space implies that $\bigcap_{n\geq 1} H_n \neq \emptyset$. If $s \in \bigcap_{n\geq 1} H_n$, then it is clear that $x \in \bigcap_{n\geq 1} L_n$ and $x \in L$, which means that the set $\bigcap_{n\geq 1} L_n$ has a nonempty intersection with every open set $L$. This implies that $\bigcap_{n\geq 1} L_n$ is dense in $S$. □

The notion of precompactness that we are about to introduce is a weaker notion than the notion of compactness formulated for general topological spaces, which can be introduced for topological metric spaces.

**Definition 11.55.** *Let $(S, d)$ be a metric space. A finite subset $\{x_1, \ldots, x_n\}$ is an* $r$-net on $(S, d)$ *if $S = \bigcup_{i=1}^{n} C(x_i, r)$.*

Observe that, for every positive number $r$ the family of open spheres $\{C(x, r) \mid x \in S\}$ is an open cover of the space $S$.

**Definition 11.56.** *A topological metric space $(S, \mathcal{O}_d)$ is* precompact *if, for every positive number $r$, the open cover $\{C(x, r) \mid x \in S\}$ contains an $r$-net $\{C(x_1, r), \ldots, C(x_n, r)\}$.*

Clearly, compactness implies precompactness.

Using the notion of an $r$-net, it is possible to give the following characterization to precompactness.

**Theorem 11.57.** *The topological metric space $(S, \mathcal{O}_d)$ is precompact if and only if for every positive number $r$ there exists an $r$-net $N_r$ on $(S, \mathcal{O}_d)$.*

*Proof.* This statement is an immediate consequence of the definition of precompactness. □

Next, we show that precompactness is inherited by subsets.

**Theorem 11.58.** *If $(S, \mathcal{O}_d)$ is a precompact topological metric space and $T \subseteq S$, then the subspace $(T, \mathcal{O}_d \restriction_T)$ is also precompact.*

*Proof.* Since $(S, \mathcal{O}_d)$ is precompact, for every $r > 0$ there exists a finite open cover $\mathcal{C}_{r/2} = \{C\left(s_i, \frac{r}{2}\right) \mid s_i \in S, 1 \leq i \leq n\}$. Let $\mathcal{C}' = \{C\left(s_{i_j}, \frac{r}{2}\right) \mid 1 \leq j \leq m\}$ be a minimal subcollection of $\mathcal{C}_{r/2}$ that consists of those open spheres that cover $T$; that is,

$$T \subseteq \bigcup \left\{ C\left(s_{i_j}, \frac{r}{2}\right) \mid 1 \leq j \leq m \right\}.$$

The minimality of $\mathcal{C}'$ implies that each set $C\left(s_{i_j}, \frac{r}{2}\right)$ contains an element $y_j$ of $T$. By Exercise 6 of Chapter 10, we have $C\left(s_{i_j}, \frac{r}{2}\right) \subseteq C(y_j, r)$ and this implies that the set $\{y_1, \ldots, y_m\}$ is an $r$-net for the set $T$. □

If the subspace $(T, \mathcal{O}_d \restriction_T)$ of $(S, \mathcal{O}_d)$ is precompact, we say that the set $T$ is *precompact*.

The next corollary shows that there is no need to require the centers of the spheres involved in the definition of the precompactness of a subspace to be located in the subspace.

**Corollary 11.59.** *Let $(S, \mathcal{O}_d)$ be a topological metric space (not necessarily precompact) and let $T$ be a subset of $S$. The subspace $(T, \mathcal{O}_d \restriction_T)$ is precompact if and only if for every positive number $r$ there exists a finite subcover $\{C(x_1, r), \ldots, C(x_n, r) \mid x_i \in S \text{ for } 1 \leq i \leq n\}$.*

*Proof.* The argument has been made in the proof of Theorem 11.58. □

The next theorem adds two further equivalent characterizations of compact metric spaces to the ones given in Theorem 11.28.

**Theorem 11.60.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. The following statements are equivalent.*

(i) *$(S, \mathcal{O}_d)$ is compact.*

(ii) *Every sequence $\boldsymbol{x} \in \mathbf{Seq}_\infty(S)$ contains a convergent subsequence.*

(iii) $(S, \mathcal{O}_d)$ *is precompact and complete.*

*Proof.* (i) implies (ii): Let $(S, \mathcal{O}_d)$ be a compact topological metric space and let $\mathbf{x}$ be a sequence in $\mathbf{Seq}_\infty(S)$. By Theorem 11.28, $(S, \mathcal{O}_d)$ has the Bolzano-Weierstrass property, so the set $\{x_n \mid n \in \mathbb{N}\}$ has an accumulation point $t$. For every $k \geq 1$, the set $\{x_n \mid n \in \mathbb{N}\} \cap C\left(t, \frac{1}{k}\right)$ contains an element $x_{n_k}$ distinct from $t$. Since $d(t, x_{n_k}) < \frac{1}{k}$ for $k \geq 1$, it follows that the subsequence $(x_{n_1}, x_{n_2}, \ldots)$ converges to $t$.

(ii) implies (iii): Suppose that every sequence $\mathbf{x} \in \mathbf{Seq}_\infty(S)$ contains a convergent subsequence and that $(S, \mathcal{O}_d)$ is not precompact. Then, there exists a positive number $r$ such that $S$ cannot be covered by any collection of open spheres of radius $r$.

Let $x_0$ be an arbitrary element of $S$. Note that $C(x_0, r) - S \neq \emptyset$ because otherwise the $C(x_0, r)$ would constitute an open cover for $S$. Let $x_1$ be an arbitrary element in $C(x_0, r) - S$. Observe that $d(x_0, x_1) \geq r$. The set $(C(x_0, r) \cup C(x_1, r)) - S$ is not empty. Thus, for any $x_2 \in (C(x_0, r) \cup C(x_1, r)) - S$, we have $d(x_0, x_2) \geq r$ and $d(x_0, x_1) \geq r$, etc. We obtain in this manner a sequence $x_0, x_1, \ldots, x_n, \ldots$ such that $d(x_i, x_j) \geq r$ when $i \neq j$. Clearly, this sequence cannot contain a convergent sequence, and this contradiction shows that the space must be precompact.

To prove that $(S, \mathcal{O}_d)$ is complete, consider a Cauchy sequence $\mathbf{x} = (x_0, x_1, \ldots, x_n, \ldots)$. By hypothesis, this sequence contains a convergent subsequence $(x_{n_0}, x_{n_1}, \ldots)$. Suppose that $\lim_{k \to \infty} x_{n_k} = l$. Since $\mathbf{x}$ is a Cauchy sequence, there is $n'_{\frac{\epsilon}{2}}$ such that $n, n_k \geq n'_{\frac{\epsilon}{2}}$ implies $d(x_n, x_{n_k}) < \frac{\epsilon}{2}$. The convergence of the subsequence $(x_{n_0}, x_{n_1}, \ldots)$ means that there exists $n''_{\frac{\epsilon}{2}}$ such that $n_k \geq n''_{\frac{\epsilon}{2}}$ implies $d(x_{n_k}, l) < \frac{\epsilon}{2}$. Choosing $n_k \geq n''_{\frac{\epsilon}{2}}$, if $n \geq n'_{\frac{\epsilon}{2}} = n_\epsilon$, we obtain

$$d(x_n, l) \leq d(x_n, x_{n_k}) + d(x_{n_k}, l) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

which proves that $\mathbf{x}$ is convergent. Consequently, $(S, \mathcal{O}_d)$ is both precompact and complete.

(iii) implies (i): Suppose that $(S, \mathcal{O}_d)$ is both precompact and complete but not compact, which means that there exists an open cover $\mathcal{C}$ of $S$ that does not contain any finite subcover.

Since $(S, \mathcal{O}_d)$ is precompact, there exists a $\frac{1}{2}$-net, $\{x_1^1, \ldots, x_{n_1}^1\}$. For each of the closed spheres $B(x_i^1, \frac{1}{2})$, $1 \leq i \leq n_1$, the trace collection $\mathcal{C}_{B(x_i^1, \frac{1}{2})}$ is an open cover. There is a closed sphere $B(x_j^1, \frac{1}{2})$ such that the open cover $\mathcal{C}_{B(x_j^1, \frac{1}{2})}$ does not contain any finite subcover of $B(x_j^1, \frac{1}{2})$ since $(S, \mathcal{O}_d)$ was assumed not to be compact. Let $z_1 = x_j^1$.

By Theorem 11.58, the closed sphere $B(z_1, \frac{1}{2})$ is precompact. Thus, there exists a $\frac{1}{2^2}$-net $\{x_1^2, \ldots, x_{n_2}^2\}$ of $B(z_1, \frac{1}{2})$. There exists a closed sphere $B(x_k^2, \frac{1}{2^2})$ such that the open cover $\mathcal{C}_{B(x_k^2, \frac{1}{2^2})}$ does not contain any finite subcover of $B(x_k^2, \frac{1}{2^2})$. Let $z_2 = x_k^2$; note that $d(z_1, z_2) \leq \frac{1}{2}$.

Thus, we construct a sequence $\mathbf{z} = (z_1, z_2, \ldots)$ such that $d(z_{n+1}, z_n) \leq \frac{1}{2^n}$ for $n \geq 1$.

Observe that

$$\begin{aligned} d(z_{n+p}, z_n) &\leq d(z_{n+p}, z_{n+p-1}) + d(z_{n+p-1}, z_{n+p-2}) + \cdots + d(z_{n+1}, z_n) \\ &\leq \frac{1}{2^{n+p-1}} \frac{1}{2^{n+p-2}} + \cdots + \frac{1}{2^n} \\ &= \frac{1}{2^{n-1}} \left(1 - \frac{1}{2^p}\right). \end{aligned}$$

Thus, the sequence $\mathbf{z}$ is a Cauchy sequence and there exists $z = \lim_{n\to\infty} z_n$, because $(S, \mathcal{O}_d)$ is complete.

Since $\mathcal{C}$ is an open cover, there exists a set $L \in \mathcal{C}$ such that $z \in L$. Let $r$ be a positive number such that $C(z, r) \subseteq L$. Let $n_0$ be such that $d(z_n, z) < \frac{r}{2}$ and $\frac{1}{2^n} \leq \frac{r}{2}$. If $x \in B(z_n, \frac{1}{2^n})$, then $d(x, z) \leq d(x, z_n) + d(z_n, z) < \frac{1}{2^n} + \frac{r}{2} \leq r$, so $B(z_n, \frac{1}{2^n}) \subseteq C(z, r) \subseteq L$. This is a contradiction because the spheres $B(z_n, \frac{1}{2^n})$ were defined such that $\mathcal{C}_{B(z_n, \frac{1}{2^n})}$ did not contain any finite subcover. Thus, $(S, \mathcal{O}_d)$ is compact. □

**Theorem 11.61.** *A subset $T$ of $(\mathbb{R}^n, \mathcal{O})$ is compact if and only if it is closed and bounded.*

*Proof.* Let $T$ be a compact set. By Corollary 11.21. $T$ is closed. Let $r$ be a positive number and let $\{C(t, r) \mid t \in T\}$ be a cover of $T$. Since $T$ is compact, there exists a finite collection $\{C(t_i, r) \mid 1 \leq i \leq p\}$ such that $T \subseteq \bigcup\{C(t_i, r) \mid 1 \leq i \leq p\}$. Therefore, if $x, y \in T$, we have $d(x, y) \leq 2 + \max\{d(t_i, t_j) \mid 1 \leq i, j \leq p\}$, which implies that $T$ is also bounded.

Conversely, suppose that $T$ is closed and bounded. The boundedness of $T$ implies the existence of a parallelepiped $[x_1, y_1] \times \cdots \times [x_n, y_n]$ that includes $T$, and we saw in Example 6.99 that this parallelepiped is compact. Since $T$ is closed, it is immediate that $T$ is compact by Theorem 6.60. □

**Corollary 11.62.** *Let $(S, \mathcal{O})$ be a compact topological space and let $f : S \longrightarrow \mathbb{R}$ be a continuous function, where $\mathbb{R}$ is equipped with the usual topology. Then, $f$ is bounded and there exist $u_0, u_1 \in S$ such that $f(u_0) = \inf_{x\in S} f(x)$ and $f(u_1) = \sup_{x\in S} f(x)$.*

*Proof.* Since $S$ is compact and $f$ is continuous, the set $f(S)$ is a compact subset of $\mathbb{R}$ and, by Theorem 11.61, is bounded and closed.

Both $\inf_{x\in S} f(x)$ and $\sup_{x\in S} f(x)$ are cluster points of $f(S)$; therefore, both belong to $f(S)$, which implies the existence of $u_0$ and $u_1$. □

**Theorem 11.63 (Heine's Theorem).** *Let $(S, \mathcal{O}_d)$ be a compact topological metric space and let $(T, \mathcal{O}_e)$ be a metric space. Every continuous function $f : S \longrightarrow T$ is uniformly continuous on $S$.*

*Proof.* Let $\mathbf{u} = (u_0, u_1, \ldots)$ and $\mathbf{v} = (v_0, v_1, \ldots)$ be two sequences in $\mathbf{Seq}_\infty(S)$ such that $\lim_{n\to\infty} d(u_n, v_n) = 0$. By Theorem 11.60, the sequence $\mathbf{u}$ contains a convergent subsequence $(u_{p_0}, u_{p_1}, \ldots)$. If $x = \lim_{n\to\infty} u_{p_n}$, then $\lim_{n\to\infty} v_{p_n} = x$. The continuity of $f$ implies that $\lim_{n\to\infty} e(f(u_{p_n}), f(v_{p_n})) = e(f(x), f(x)) = 0$, so $f$ is uniformly continuous by Theorem 11.13. □

## 11.7 Contractions and Fixed Points

**Definition 11.64.** *Let $(S, d)$ and $(T, d')$ be two metric spaces. A function $f : S \longrightarrow T$ is a* similarity *if there exists a number $r > 0$ for which $d'(f(x), f(y)) = rd(x, y)$ for every $x, y \in S$. If the two metric spaces coincide, we refer to $f$ as a* self-similarity of $(S, d)$.

*The number $r$ is called the* ratio *of the similarity $f$ and is denoted by $ratio(f)$.*

*An* isometry *is a similarity of ratio* 1. *If an isometry exists between the metric spaces $(S, d)$ and $(T, d')$, then we say that these spaces are* isometric.

*If there exists $r > 0$ such that $d'(f(x), f(y)) \leq rd(x, y)$ for all $x, y \in S$, then we say that $f$ is a* Lipschitz function. *Furthermore, if this inequality is satisfied for a number $r < 1$, then $f$ is a* contraction.

*Example 11.65.* Let $(\mathbb{R}, d)$ be the metric space defined by $d(x, y) = |x - y|$. Any linear mapping (that is, any mapping of the form $f(x) = ax+b$ for $x \in \mathbb{R}$) is a similarity having ratio $a$.

**Theorem 11.66.** *Let $(S, \mathcal{O}_d)$ and $(T, \mathcal{O}_{d'})$ be two metric spaces. Every Lipschitz function $f : S \longrightarrow T$ is continuous.*

*Proof.* Let $\epsilon$ be a positive number. Define $\delta = \frac{\epsilon}{k}$. If $z \in f(C(x, \delta))$, there exists $y \in C(x, \delta)$ such that $z = f(y)$. This implies $e(f(x), z) = e(f(x), f(y)) < kd(x, y) < k\delta = \epsilon$, so $z \in C(f(x), \epsilon)$. Thus, $f(C(x, \delta)) \subseteq C(f(x), \epsilon)$, so $f$ is continuous. □

Theorem 11.66 implies that every similarity is continuous.

Let $f : S \longrightarrow S$ be a function. We define inductively the functions $f^{(n)} : S \longrightarrow S$ for $n \in \mathbb{N}$ by

$$f^{(0)}(x) = x$$

and

$$f^{(n+1)}(x) = f(f^{(n)}(x))$$

for $x \in S$. The function $f^{(n)}$ is the $n^{\text{th}}$ *iteration of the function $f$.*

*Example 11.67.* Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be the function defined by $f(x) = ax + b$ for $x \in \mathbb{R}$, where $a, b \in \mathbb{R}$ and $a \neq 1$. We have

$$\begin{aligned} f^{(0)}(x) &= x, \\ f^{(1)}(x) &= ax + b, \\ f^{(2)}(x) &= a^2x + ab, \\ &\vdots \\ f^{(n)}(x) &= a^n x + \frac{a^n - 1}{a - 1} \cdot b, \\ &\vdots \end{aligned}$$

for $x \in \mathbb{R}$.

**Definition 11.68.** *Let $f : S \longrightarrow S$ be a function. A* fixed point of $f$ *is a member $x$ of the set $S$ that satisfies the equality $f(x) = x$.*

*Example 11.69.* The function $f$ defined in Example 11.67 has the fixed point $x_0 = \frac{b}{1-a}$.

**Theorem 11.70 (Banach Fixed Point Theorem).** *Let $(S, \mathcal{O}_d)$ be a complete topological metric space and let $f : S \longrightarrow S$ be a contraction on $S$. Then, there exists a unique fixed point $u \in S$ for $f$, and for any $x \in S$ we have $\lim_{n\to\infty} f^{(n)}(x) = u$.*

*Proof.* Since $f$ is a contraction, there exists a positive number $r$, $r < 1$, such that $d(f(x), f(y)) \leq rd(x, y)$ for $x, y \in S$. Note that each such function has at most one fixed point. Indeed, suppose that we have both $u = f(u)$ and $v = f(v)$ and $u \neq v$, so $d(u, v) > 0$. Then, $d(f(u), f(v)) = d(u, v) \leq rd(u, v)$, which is absurd because $r < 1$.

The sequence $\mathbf{s} = (x, f(x), \ldots, f^{(n)}(x), \ldots)$ is a Cauchy sequence. Indeed, observe that

$$d(f^{(n)}(x), f^{(n+1)}(x)) \leq rd(f^{(n-1)}(x), f^{(n)}(x)) \leq \cdots \leq r^n d(x, f(x)).$$

For $n \leq p$, this implies

$$\begin{aligned} d(f^{(n)}(x), f^{(p)}(x)) &\leq d(f^{(n)}(x), f^{(n+1)}(x)) + d(f^{(n+1)}(x), f^{(n+2)}(x)) + \\ &\quad \cdots + d(f^{(p-1)}(x), f^{(p)}(x)) \\ &\leq r^n d(x, f(x) + \cdots + r^{p-1} d(x, f(x)) \\ &\leq \frac{r^n}{1 - r} d(x, f(x)), \end{aligned}$$

which shows that the sequence $\mathbf{s}$ is indeed a Cauchy sequence. Since $(S, \mathcal{O}_d)$ is complete, there exists $u \in S$ such that $u = \lim_{n\to\infty} f^{(n)}(x)$. The continuity of $f$ implies

$$u = \lim_{n\to\infty} f^{(n+1)}(x) = \lim_{n\to\infty} f(f^{(n)}(x)) = f(u),$$

so $u$ is a fixed point of $f$. Since $d(f^{(n)}(x), f^{(p)}(x)) \leq \frac{r^n}{1-r} d(x, f(x))$, we have

$$\lim_{p \to \infty} d(f^{(n)}(x), f^{(p)}(x)) = d(f^{(n)}(x), u) \leq \frac{r^n}{1-r} d(x, f(x))$$

for $n \in \mathbb{N}$. □

**The Hausdorff Metric Hyperspace of Compact Subsets**

**Lemma 11.71.** *Let $(S, d)$ be a metric space and let $U$ and $V$ be two subsets of $S$. If $r \in \mathbb{R}_{\geq 0}$ is such that $U \subseteq C(V, r)$ and $V \subseteq C(U, r)$, then we have $|d(x, U) - d(x, V)| \leq r$ for every $x \in S$.*

*Proof.* Since $U \subseteq C(V, r)$, for every $u \in U$ there is $v \in V$ such that $d(u, v) < r$. Therefore, by the triangular inequality, it follows that for every $u \in U$ there is $v \in V$ such that $d(x, u) < d(x, v) + r$, so $d(x, U) < d(x, v) + r$. Consequently, $d(x, U) \leq d(x, V) + r$. In a similar manner, we can show that $V \subseteq C(U, r)$ implies $d(x, V) \leq d(x, U) + r$. Thus, $|d(x, U) - d(x, V)| \leq r$ for every $x \in S$. □

Let $(S, \mathcal{O}_d)$ be a topological metric space. Denote by $\mathcal{K}(S, \mathcal{O}_d)$ the collection of all nonempty, compact subsets of $(S, \mathcal{O}_d)$, and define the mapping $\delta : \mathcal{K}(S, \mathcal{O}_d)^2 \longrightarrow \mathbb{R}_{\geq 0}$ by

$$\delta(U, V) = \inf\{r \in \mathbb{R}_{\geq 0} \mid U \subseteq C(V, r) \text{ and } V \subseteq C(U, r)\}$$

for $U, V \in \mathcal{K}(S, \mathcal{O}_d)$.

**Lemma 11.72.** *Let $U$ and $V$ be two compact subsets of a topological metric space $(S, \mathcal{O}_d)$. We have*

$$\sup_{x \in S} |d(x, U) - d(x, V)| = \max\left\{\sup_{x \in V} d(x, U), \sup_{x \in U} d(x, V)\right\}.$$

*Proof.* Let $x \in S$. There is $v_0 \in V$ such that $d(x, v_0) = d(x, V)$ because $V$ is a compact set. Then, the compactness of $U$ implies that there is $u_0 \in U$ such that $d(u_0, v_0) = d(v_0, U)$. We have

$$\begin{aligned} d(x, U) - d(x, V) &= d(x, U) - d(x, v_0) \\ &\leq d(x, u_0) - d(x, v_0) \\ &\leq d(u_0, v_0) \leq \sup_{x \in V} d(U, x). \end{aligned}$$

Similarly, $d(x, U) - d(x, V) \leq \sup_{x \in U} d(x, V)$, which implies

$$\sup_{x \in S} |d(x, U) - d(x, V)| \leq \max\left\{\sup_{x \in V} d(x, U), \sup_{x \in U} d(x, V)\right\}.$$

On the other hand, since $U \subseteq S$, we have

$$\sup_{x \in S} |d(x, U) - d(x, V)| \geq \sup_{x \in U} |d(x, U) - d(x, V)| = \sup_{x \in U} d(x, V)$$

and, similarly, $\sup_{x \in S} |d(x, U) - d(x, V)| \geq \sup_{x \in V} d(x, U)$, and these inequalities prove that

$$\sup_{x \in S} |d(x, U) - d(x, V)| \geq \max \left\{ \sup_{x \in V} d(x, U), \sup_{x \in U} d(x, V) \right\},$$

which concludes the argument. □

An equivalent useful definition of $\delta$ is given in the next theorem.

**Theorem 11.73.** *Let $(S, d)$ be a metric space and let $U$ and $V$ be two compact subsets of $S$. We have the equality*

$$\delta(U, V) = \sup_{x \in S} |d(x, U) - d(x, V)|.$$

*Proof.* Observe that we have both $U \subseteq C(V, \sup_{x \in U} d(x, V))$ and $V \subseteq C(U, \sup_{x \in V} d(x, U))$. Therefore, we have

$$\delta(U, V) \leq \max\{\sup_{x \in V} d(x, U), \sup_{x \in U} d(x, V)\}.$$

Combining this observation with Lemma 11.72 yields the desired equality. □

**Theorem 11.74.** *Let $(S, \mathcal{O}_d)$ be a complete topological metric space. The mapping $\delta : \mathcal{K}(S, \mathcal{O}_d)^2 \longrightarrow \mathbb{R}_{\geq 0}$ is a metric on $\mathcal{K}(S, \mathcal{O}_d)$.*

*Proof.* It is clear that $\delta(U, U) \geq 0$ and that $\delta(U, V) = \delta(V, U)$ for every $U, V \in \mathcal{K}(S, \mathcal{O}_d)$. Suppose that $\delta(U, V) = 0$; that is, $d(x, U) = d(x, V)$ for every $x \in S$. If $x \in U$, then $d(x, U) = 0$, so $d(x, V) = 0$. Since $V$ is closed, by Part (ii) of Theorem 11.16, we have $x \in V$, so $U \subseteq V$. The reverse inclusion can be shown in a similar manner.

To prove the triangular inequality, let $U, V, W \in \mathcal{K}(S, \mathcal{O}_d)$. Since

$$|d(x, U) - d(x, V)| \leq |d(x, U) - d(x, V)| + |d(x, V) - d(x, W)|,$$

for every $x \in S$, we have

$$\begin{aligned} \sup_{x \in S} |d(x, U) - d(x, V)| &\leq \sup_{x \in S} (|d(x, U) - d(x, V)| + |d(x, V) - d(x, W)|) \\ &\leq \sup_{x \in S} |d(x, U) - d(x, V)| + \sup_{x \in S} |d(x, V) - d(x, W)|, \end{aligned}$$

which implies the triangular inequality

$$\delta(U, V) \leq \delta(U, W) + \delta(W, V).$$

□

The metric $\delta$ is known as the *Hausdorff metric*, and the metric space $(\mathcal{K}(S, \mathcal{O}_d), \delta)$ is known as the Hausdorff metric hyperspace of $(S, \mathcal{O}_d)$.

**Theorem 11.75.** *If $(S, \mathcal{O}_d)$ is a complete topological metric space, then so is the Hausdorff metric hyperspace $(\mathcal{K}(S, \mathcal{O}_d), \delta)$.*

*Proof.* Let $\mathbf{U} = (U_0, U_1, \ldots)$ be a Cauchy sequence in $(\mathcal{K}(S, \mathcal{O}_d), \delta)$ and let $U = \mathbf{K}(\bigcup_{n\in\mathbb{N}} U_n)$. It is clear that $U$ consists of those elements $x$ of $S$ such that $x = \lim_{n\to\infty} x_n$ for some sequence $\mathbf{x} = (x_0, x_1, \ldots)$, where $x_n \in U_n$ for $n \in \mathbb{N}$.

The set $U$ is precompact. Indeed, let $\epsilon > 0$ and let $n_0$ be such that $\delta(U_n, U_{n_0}) \leq \epsilon$ for $n \geq n_0$. Let $N$ be an $\epsilon$-net for the compact set $H = \bigcup_{n\leq n_0} U_n$. Clearly, $H \subseteq C(N, \epsilon)$. Since $\delta(U_n, U_{n_0}) \leq \epsilon$, it follows that $U \subseteq C(H, \epsilon)$, so $U \subseteq C(N, 2\epsilon)$. This shows that $U$ is precompact. Since $U$ is closed in the complete space $(S, \mathcal{O}_d)$, it follows that $U$ is compact.

Let $\epsilon$ be a positive number. Since $\mathbf{U}$ is a Cauchy sequence, there exists $n_{\frac{\epsilon}{2}}$ such that $m, n \geq n_{\frac{\epsilon}{2}}$ implies $\delta(U_m, U_n) < \frac{\epsilon}{2}$; that is, $\sup_{s\in S} |d(s, U_m) - d(s, U_n)| < \frac{\epsilon}{2}$. In particular, if $x_m \in U_m$, then $d(x_m, U_n) = \inf_{y\in U_m} d(x, y) < \frac{\epsilon}{2}$, so there exists $y \in U_n$ such that $d(x_m, y) < \frac{\epsilon}{2}$.

For $x \in U$, there exists a sequence $\mathbf{x} = (x_0, x_1, \ldots)$ such that $x_n \in U_n$ for $n \in \mathbb{N}$ and $\lim_{n\to\infty} x_n = x$. Therefore, there exists a number $n'_{\frac{\epsilon}{2}}$ such that $p \geq n'_{\frac{\epsilon}{2}}$ implies $d(x, x_p) < \frac{\epsilon}{2}$. This implies $d(x, y) \leq d(x, x_p) + d(x_p, y) \leq \epsilon$ if $n \geq \max\{n_{\frac{\epsilon}{2}}, n'_{\frac{\epsilon}{2}}\}$, and therefore $U \subseteq C(U_n, \epsilon)$.

Let $y \in U_n$. Since $\mathbf{U}$ is a Cauchy sequence, there exists a subsequence $\mathbf{U}' = (U_{k_0}, U_{k_1}, \ldots)$ of $\mathbf{U}$ such that $k_0 = q$ and $\delta(U_{k_j}, U_n) < 2^j\epsilon$ for all $n \geq k_j$.

Define the sequence $\mathbf{z} = (z_0, z_1, \ldots)$ by choosing $z_k$ arbitrarily for $k < q$, $z_q = y$, and $z_k \in U_k$ for $k_j < k < k_{j+1}$ such that $d(z_k, z_{k_j}) < 2^{-j}\epsilon$. The sequence $\mathbf{z}$ is a Cauchy sequence in $S$, so there exists $z = \lim_{k\to\infty} z_k$ and $z \in U$. Since $d(y, z) = \lim_{k\to\infty} d(y, z_k) < \epsilon$, it follows that $y \in C(U, \epsilon)$. Therefore, $\delta(U, U_n) < \epsilon$, which proves that $\lim_{n\to\infty} U_n = U$. We conclude that $(\mathcal{K}(S, \mathcal{O}_d), \delta)$ is complete. □

## 11.8 Measures in Metric Spaces

In this section, we discuss the interaction between metrics and measures defined on metric spaces.

**Definition 11.76.** *Let $(S, d)$ be a metric space. A* Carathéodory outer measure *on $(S, d)$ is an outer measure on $S$, $\mu : \mathcal{P}(S) \longrightarrow \hat{\mathbb{R}}_{\geq 0}$ such that, for every two sets $U$ and $V$ of the topological space $(S, \mathcal{O}_d)$ such that $d(U, V) > 0$, we have $\mu(U \cup V) = \mu(U) + \mu(V)$.*

*Example 11.77.* The Lebesgue outer measure introduced in Example 6.130 is a Carathéodory outer measure.

Indeed, let $U$ and $V$ be two disjoint subsets of $\mathbb{R}^n$ such that $d_2(U, V) > 0$ and let $\mathcal{D}$ be the family of $n$-dimensional intervals that covers $U \cup V$. Suppose that the diameter of each of these intervals is less than $r$. If $D = \bigcup \mathcal{D}$, we

have $D = D_U \cup D_V \cup D'$, where $U \subseteq D_U$, $V \subseteq D_V$, and $D' \cap (U \cup V) = \emptyset$. Since $vol(D_U) + vol(D_V) \leq vol(D)$, we have $\mu(U) + \mu(V) \leq \mu(U \cup V)$, which implies that $\mu$ is a Carathéodory outer measure.

**Theorem 11.78.** *Let $(S, d)$ be a metric space. The outer measure $\mu$ on $S$ is a Carathéodory outer measure if and only if every closed set of $(S, \mathcal{O}_d)$ is $\mu$-measurable.*

*Proof.* Suppose that every closed set is $\mu$-measurable, and let $U$ and $V$ be two subsets of $S$ such that $d(U, V) > 0$. Consider the closed set $\mathbf{K}(C(U, r))$, where $r = \frac{d(u,v)}{2}$. Since this is a $\mu$-measurable set, we have

$$\mu(U \cup V) = \mu((U \cup V) \cap C(U, r)) + \mu((U \cup V) \cap \mathbf{K}(C(U, r))) = \mu(U) + \mu(V),$$

so $\mu$ is a Carathéodory outer measure.

Conversely, suppose that $\mu$ is a Carathéodory outer measure; that is, $d(U, V) > 0$ implies $\mu(U \cup V) = \mu(U) + \mu(V)$.

Let $U$ be an open set, $L$ be a subset of $U$, and $L_1, L_2, \ldots$ be a sequence of sets defined by

$$L_n = \left\{ t \in L \mid d(t, \mathbf{K}(U)) \geq \frac{1}{n} \right\}$$

for $n \geq 1$. Note that $L_1, L_2, \ldots$ is an increasing sequence of sets, so the sequence $\mu(L_1), \mu(L_2), \ldots$ is increasing. Therefore, $\lim_{n\to\infty} \mu(L_i)$ exists and $\lim_{n\to\infty} \mu(L_i) \leq \mu(L)$. We claim that $\lim_{n\to\infty} \mu(L_i) = \mu(L)$.

Since every set $L_n$ is a subset of $L$, it follows that $\bigcup_{n\geq 1} L_n \subseteq L$. Let $t \in L \subseteq U$. Since $U$ is an open set, there exists $\epsilon > 0$ such that $C(t, \epsilon) \subseteq U$, so $d(t, \mathbf{K}(U)) \geq \frac{1}{n}$ if $n > \frac{1}{\epsilon}$. Thus, for sufficiently large values of $n$, we have $t \in L_n$, so $L \subseteq \bigcup_{n\geq 1} L_n$. This shows that $L = \bigcup_{n\geq 1} L_n$.

Consider the sequence of sets $M_n = L_{n+1} - L_n$ for $n \geq 1$. Clearly, we can write

$$L = L_{2n} \cup \bigcup_{k=2n}^{\infty} M_k = L_{2n} \cup \bigcup_{p=n}^{\infty} M_{2p} \cup \bigcup_{p=n}^{\infty} M_{2p+1},$$

so

$$\mu(L) \leq \mu(L_{2n}) + \sum_{p=n}^{\infty} \mu(M_{2p}) + \sum_{p=n}^{\infty} \mu(M_{2p+1}).$$

If both series $\sum_{p=1}^{\infty} \mu(M_{2p})$ and $\sum_{p=1}^{\infty} \mu(M_{2p+1})$ are convergent, then

$$\lim_{n\to\infty} \sum_{p=n}^{\infty} \mu(M_{2p}) = 0 \text{ and } \lim_{n\to\infty} \sum_{p=n}^{\infty} \mu(M_{2p+1}) = 0,$$

and so $\mu(L) \leq \lim_{n\to\infty} \mu(L_{2n})$.

If the series $\sum_{p=n}^{\infty} \mu(M_{2p})$ is divergent, let $t \in M_{2p} \subseteq L_{2p+1}$. If $z \in \mathbf{K}(U)$, then $d(t, z) \geq \frac{1}{2p+1}$ by the definition of $L_{2p+1}$. Let $y \in M_{2p+2} \subseteq L_{2p+3}$. We have

$$\frac{1}{2p+2} > d(y,z) > \frac{1}{2p+3},$$

so

$$d(t,y) \geq t(t,z) - d(y,z) \geq \frac{1}{2p+1} - \frac{1}{2p+2},$$

which means that $d(M_{2p}, M_{2p+2}) > 0$ for $p \geq 1$. Since $\mu$ is a Carathéodory outer measure, we have

$$\sum_{p=1}^{n} \mu(M_{2p}) = \mu\left(\bigcup_{p=1}^{n} M_{2p}\right) \leq \mu(L_{2n}).$$

This implies $\lim_{n\to\infty} \mu(L_n) = \lim_{n\to\infty} \mu(L_{2n}) = \infty$, so we have in all cases $\lim_{n\to\infty} \mu(A_n) = \mu(L)$.

Let $F$ be a closed set in $(S, \mathcal{O}_d)$ and let $V$ be an arbitrary set. The set $V \cup \mathbf{K}(F)$ is contained in the set $\mathbf{K}(F) = F$, so, by the previous argument, there exists a sequence of sets $L_n$ such that $d(L_n, F) \geq \frac{1}{n}$ for each $n$ and $\lim_{n\to\infty} \mu(L_n) = \mu(V \cap \mathbf{K}(F))$. Consequently, $\mu(V) \geq \mu((V \cap F) \cup L_n) = \mu(V \cup F) + \mu(L_n)$. Taking the limit, we obtain $\mu(V) \geq \mu(V \cap F) + \mu(V \cap \mathbf{K}(F))$, which proves that $F$ is $\mu$-measurable. □

**Corollary 11.79.** *Let $(S, d)$ be a metric space. Every Borel subset of $S$ is $\mu$-measurable, where $\mu$ is a Carathéodory outer measure on $S$.*

*Proof.* Since every closed set is $\mu$-measurable relative to a Carathéodory outer measure, it follows that every Borel set is $\mu$-measurable with respect to such a measure. □

Thus, we can conclude that every Borel subset of $S$ is Lebesgue measurable.

Let $(S, d)$ be a metric space and let $\mathcal{C}$ be a countable collection of subsets of $S$. Define

$$\mathcal{C}_r = \{C \in \mathcal{C} \mid \mathit{diam}(C) < r\},$$

and assume that for every $x \in S$ and $r > 0$ there exists $C \in \mathcal{C}_r$ such that $x \in C$. Thus, the collection $\mathcal{C}_r$ is a sequential cover for $S$, and for every function $f : \mathcal{C} \to \hat{\mathbb{R}}_{\geq 0}$ we can construct an outer measure $\mu_{f,r}$ using Method I (the method described in Theorem 6.127). This construction yields an outer measure that is not necessarily a Carathéodory outer measure.

By Corollary 6.129, when $r$ decreases, $\mu_{f,r}$ increases. This allows us to define

$$\hat{\mu}_f = \lim_{r\to 0} \mu_{f,r}.$$

We shall prove that the measure $\hat{\mu}_f$ is a Carathéodory outer measure.

Since each measure $\mu_{f,r}$ is an outer measure, it follows immediately that $\hat{\mu}_f$ is an outer measure.

**Theorem 11.80.** *Let $(S, d)$ be a metric space, $\mathcal{C}$ be a countable collection of subsets of $S$, and $f : \mathcal{C} \to \hat{\mathbb{R}}_{\geq 0}$. The measure $\hat{\mu}_f$ is a Carathéodory outer measure.*

*Proof.* Let $U$ and $V$ be two subsets of $S$ such that $d(U,V) > 0$. We need to show only that $\hat{\mu}_f(U \cup V) \geq \hat{\mu}_f(U) + \hat{\mu}_f(V)$.

Choose $r$ such that $0 < r < d(U,V)$, and let $\mathcal{D}$ be an open cover of $U \cup V$ that consists of sets of $\mathcal{C}_r$. Each set of $\mathcal{D}$ can intersect at most one of the set $U$ and $V$. This observation allows us to write $\mathcal{D}$ as a disjoint union of two collections, $\mathcal{D} = \mathcal{D}_U \cup \mathcal{D}_V$, where $\mathcal{D}_U$ is an open cover for $U$ and $\mathcal{D}_V$ is an open cover for $V$. Then,

$$\begin{aligned}\sum\{f(D) \mid D \in \mathcal{D}\} &= \sum\{f(D) \mid D \in \mathcal{D}_U\} + \sum\{f(D) \mid D \in \mathcal{D}_V\}\\ &\geq \mu_{f,r}(U) + \mu_{f,r}(V).\end{aligned}$$

This implies $\mu_{f,r}(U \cup V) \geq \mu_{f,r}(U) + \mu_{f,r}(V)$, which yields $\hat{\mu}_f(U \cup V) \geq \hat{\mu}_f(U) + \hat{\mu}_f(V)$ by taking the limit for $r \to 0$. □

The construction of the Carathéodory outer measure $\hat{\mu}_f$ described earlier is knows as *Munroe's Method II* or simply *Method II* (see [100, 44]).

## 11.9 Embeddings of Metric Spaces

Searching in multimedia databases and visualization of the objects of such databases is facilitated by representing objects in a $k$-dimensional space, as observed in [49]. In general, the starting point is the matrix of distances between objects, and the aim of the representation is to preserve as much as possible the distances between objects.

**Definition 11.81.** *Let $(S,d)$ and $(S',d'))$ be two metric spaces. An* embedding *of $(S,d)$ in $(S',d')$ is a function $f : S \longrightarrow S'$. The embedding $f$ is an* isometry *if $d'(f(x),f(y)) = cd(x,y)$ for some positive constant number $c$. If $f$ is an isometry, we refer to it as an* isometric embedding.

*If an isometric embedding $f : S \longrightarrow S'$ exists, then we say that $(S,d)$ is* isometrically embedded in $(S,d)$.

Note that an isometry is an injective function for if $f(x) = f(y)$, then $d'(f(x),f(y)) = 0$, which implies $d(x,y) = 0$. This, in turn, implies $x = y$.

*Example 11.82.* Let $S$ be a set that consists of four objects, $S = \{o_1, o_2, o_3, o_4\}$, that are equidistant in the metric space $(S,d)$; in other words, we assume that $d(o_i,o_j) = k$ for every pair of distinct objects $(o_i,o_j)$.

The subset $U = \{o_1,o_2\}$ of $S$ can be isometrically embedded in $\mathbb{R}^1$; the isometry $h : U \longrightarrow \mathbb{R}^1$ is defined by $h(o_1) = (0)$ and $h(o_2) = (k)$.

For the subset $\{o_1,o_2,o_3\}$ define the embedding $f : \{o_1,o_2,o_3\} \longrightarrow \mathbb{R}^2$ by

$$\begin{aligned}f(o_1) &= (0,0),\\ f(o_2) &= (k,0),\\ f(o_3) &= (c_1,c_2),\end{aligned}$$

subject to the conditions

$$\begin{aligned} c_1^2 + c_2^2 &= k^2, \\ (c_1 - k)^2 + c_2^2 &= k^2. \end{aligned}$$

These equalities yield $c_1 = \frac{k}{2}$ and $c_2^2 = \frac{3k^2}{4}$. Choosing the positive solution of the last equality yields $f(o_3) = (\frac{k}{2}, \frac{k\sqrt{3}}{2})$.

To obtain an isometric embedding $g$ of $S$ in $\mathbb{R}^3$, we seek the mapping $g : S \longrightarrow \mathbb{R}^3$ as

$$\begin{aligned} g(o_1) &= (0, 0, 0), \\ g(o_2) &= (k, 0, 0), \\ g(o_3) &= \left( \frac{k}{2}, \frac{k\sqrt{3}}{2}, 0 \right), \\ g(o_4) &= (e_1, e_2, e_3), \end{aligned}$$

where

$$\begin{aligned} e_1^2 + e_2^2 + e_3^2 &= k^2, \\ (e_1 - k)^2 + e_2^2 + e_3^2 &= k^2, \\ \left( e_1 - \frac{k}{2} \right)^2 + \left( e_2 - \frac{k\sqrt{3}}{2} \right)^2 + e_3^2 &= k^2. \end{aligned}$$

The first two equalities imply $e_1 = \frac{k}{2}$; this, in turn, yields

$$\begin{aligned} e_2^2 + e_3^2 &= \frac{3k^2}{4}, \\ \left( e_2 - \frac{k\sqrt{3}}{2} \right)^2 + e_3^2 &= k^2. \end{aligned}$$

Subtracting these equalities, one gets $e_2 = \frac{k\sqrt{3}}{6}$. Finally, we have $e_3^2 = \frac{2k^2}{3}$. Choosing the positive solution, we obtain the embedding

$$\begin{aligned} g(o_1) &= (0, 0, 0), \\ g(o_2) &= (k, 0, 0), \\ g(o_3) &= \left( \frac{k}{2}, \frac{k\sqrt{3}}{2}, 0 \right), \\ g(o_4) &= \left( \frac{k}{2}, \frac{k\sqrt{3}}{6}, \frac{k\sqrt{6}}{3} \right). \end{aligned}$$

*Example 11.83.* Let $(S, d)$ be a finite metric space such that $|S| = n$. We show that there exists an isometric embedding of $(S, d)$ into $(\mathbb{R}^{n-1}, d_\infty)$, where $d_\infty$ was defined by Equality (10.7).

Indeed, suppose that $S = \{x_1, \ldots, x_n\}$, and define $f : S \longrightarrow \mathbb{R}^{n-1}$ as

$$f(x_i) = (d(x_1, x_i), \ldots, d(x_{n-1}, x_i))$$

for $1 \leq i \leq n$. We prove that $d(x_i, x_j) = d_\infty(f(x_i), f(x_j))$ for $1 \leq i, j \leq n$, which will imply that $f$ is an isometry with $c = 1$.

By the definition of $d_\infty$, we have

$$d_\infty(f(x_i), f(x_j)) = \max_{1 \leq k \leq n-1} |d(x_k, x_i) - d(x_k, x_j)|.$$

Note that for every $k$ we have $|d(x_k, x_i) - d(x_k, x_j)| \leq d(x_i, x_j)$ (see Exercise 1). Moreover, for $k = i$, we have $|d(x_i, x_i) - d(x_i, x_j)| = d(x_i, x_j)$, so $\max_{1 \leq k \leq n-1} |d(x_k, x_i) - d(x_k, x_j)| = d(x_i, x_j)$. The isometry whose existence was established in this example is known as the *Fréchet isometry* and was obtained in [54].

*Example 11.84.* We now prove the existence of an isometry between the metric spaces $(\mathbb{R}^2, d_\infty)$ and $(\mathbb{R}^2, d_1)$.

Consider the function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ defined by

$$f(u, v) = \left(\frac{u - v}{2}, \frac{u + v}{2}\right)$$

for $(u, v) \in \mathbb{R}^2$.

Since $\max\{a, b\} = \frac{1}{2}(|a - b| + |a + b|)$ for every $a, b \in \mathbb{R}$, it is easy to see that

$$\max\{|u - u'|, |v - v'|\} = \frac{1}{2}\Big|u - u' - (v - v')\Big| + \Big|u - u' + (v - v')\Big|,$$

which is equivalent to

$$d_\infty((u, v), (u', v')) = d_1\left(\left(\frac{u - v}{2}, \frac{u + v}{2}\right), \left(\frac{u' - v'}{2}, \frac{u' + v'}{2}\right)\right).$$

The last equality shows that $f$ is an isometry between $(\mathbb{R}^2, d_\infty)$ and $(\mathbb{R}^2, d_1)$.

## Exercises and Supplements

1. Prove that any subset $U$ of a topological metric space $(S, \mathcal{O}_d)$ that has a finite diameter is included in a closed sphere $B(x, diam(U))$ for some $x \in U$.
2. Let $d_u$ be the metric defined in Exercise 7 of Chapter 10, where $S = \mathbb{R}^2$ and $d$ is the usual Euclidean metric on $\mathbb{R}^2$.

a) Prove that if $x \neq u$, the set $\{x\}$ is open.
b) Prove that the topological metric space $(\mathbb{R}^2, \mathcal{O}_{d_u})$ is not separable.

3. Let $(S, \mathcal{O}_d)$ be a topological metric space.
   a) Prove that, for every $s \in S$ and every positive number $r$, we have $\mathbf{K}(C(x,r)) \subseteq B(x,r)$.
   b) Give an example of a topological metric space where the inclusion $\mathbf{K}(C(x,r)) \subseteq B(x,r)$ can be strict.

   **Solution** for Part (b): Let $(\mathbf{Seq}_\infty(\{0,1\}), d_\phi)$ be the ultrametric space, where the ultrametric $d_\phi$ was introduced in Supplement 46 of Chapter 10 and $\phi(\mathbf{u}) = \frac{1}{|\mathbf{u}|}$ for $\mathbf{u} \in \mathbf{Seq}(\{0,1\})$.

   It is clear that $B(\mathbf{x}, 1) = \mathbf{Seq}_\infty(\{0,1\})$ because $d_\phi(\mathbf{x}, \mathbf{y}) \leq 1$ for every $\mathbf{x}, \mathbf{y} \in \mathbf{Seq}_\infty(\{0,1\})$. On the other hand, $C(\mathbf{x}, 1)$ contains those sequences $\mathbf{y}$ that have a non-null longest common prefix with $\mathbf{x}$; the first symbol of each such sequence is the same as the first symbol $s$ of $\mathbf{x}$.

   Since $d_\phi$ is an ultrametric, the open sphere $C(\mathbf{x}, 1)$ is also closed, so $C(\mathbf{x}, 1) = \mathbf{K}(C(\mathbf{x}, 1))$. Let $s'$ be a symbol in $S$ distinct from $s$ and let $\mathbf{z} = (s', s', \ldots) \in \mathbf{Seq}_\infty(S)$. Note that $d_\phi(\mathbf{x}, \mathbf{z}) = 1$, so $\mathbf{z} \notin C(\mathbf{x}, 1) = \mathbf{K}(C(\mathbf{x}, 1))$. Thus, we have $\mathbf{K}(C(\mathbf{x}, 1) \subset B(\mathbf{x}, 1)$.

4. Consider the ultrametric space $(\mathbf{Seq}_\infty(\{0,1\}), d_\phi)$, where $\phi(\mathbf{u}) = \frac{1}{2^{|\mathbf{u}|}}$ for $\mathbf{u} \in \mathbf{Seq}(\{0,1\})$. For $\mathbf{u} \in \mathbf{Seq}(\{0,1\})$, let $P_\mathbf{u} = \{\mathbf{ut} \mid \mathbf{t} \in \mathbf{Seq}_\infty(\{0,1\})\}$ be the set that consists of all infinite sequences that begin with $\mathbf{u}$. Prove that $\{P_\mathbf{u} \mid \mathbf{u} \in \mathbf{Seq}(\{0,1\})\}$ is a basis for the topological ultrametric space defined above.
5. Let $(S, \mathcal{O}_d)$ be a topological metric space and let $U$ and $V$ be two subsets of $S$ such that $\delta(U, V) \leq r$. Prove that if $\mathcal{D} = \{D_i \mid i \in I\}$ is a cover for $V$, then the collection $\mathcal{D}' = \{C(D_i, r) \mid i \in I\}$ is a cover for $U$.
6. Prove that if $N_r$ is an $r$-net for a subset $T$ of a topological metric space $(S, \mathcal{O}_d)$, then $T \subseteq C(N_r, r)$.

   **Solution:** Suppose $N_r = \{y_i \mid 1 \leq i \leq n\}$, so $T \subseteq \bigcup_{i=1}^n C(y_i, r)$. Thus, for each $t \in T$ there is $y_j \in N_r$ such that $d(y_j, t) < r$. This implies that $d(t, N_r) = \inf\{d(t, y) \mid y \in N_r\} < r$, so $t \in C(N_r, r)$.
7. Prove that if $N_r$ is an $r$-net for each of the sets of a collection of subsets of a metric space $(S, d)$, then $N_r$ is an $r$-net for $\bigcup \mathcal{C}$.
8. Let $U$ and $V$ be two subsets of a metric space $(S, d)$ such that $\delta(U, V) \leq c$. Prove that every $r$-net for $V$ is an $(r + c)$-net for $U$.
9. Let $\{U_i \mid i \in I\}$ be a collection of pairwise disjoint open subsets of the topological metric space $(\mathbb{R}^n, \mathcal{O}_{d_2})$ such that for each $i \in I$ there exists $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^n$ such that $B(\mathbf{x}_i, ar) \subseteq V_i \subseteq B(\mathbf{y}_i, br)$. Then, for any $B(\mathbf{u}, r)$, we have $|\{V_i \mid \mathbf{K}(V_i) \cap B(\mathbf{u}, r) \neq \emptyset\}| \leq \left(\frac{1+2b}{a}\right)^n$.

   **Solution:** Suppose that $\mathbf{K}(V_i) \cap B(\mathbf{u}, r) \neq \emptyset$. Then $\mathbf{K}(V_i) \subseteq B(\mathbf{u}, r + 2br)$. Recall that the volume of a sphere of radius $r$ in $\mathbb{R}^n$ is $V_n(r) = \frac{\pi^{\frac{n}{2}} r^n}{\Gamma\left(\frac{n}{2}+1\right)}$ as shown in Section C.4. If $m = |\{V_i \mid \mathbf{K}(V_i) \cap B(\mathbf{u}, r) \neq \emptyset\}|$, the total volume of the spheres $B(\mathbf{x}_i, ar)$ is smaller than the volume of the sphere $B(\mathbf{u}, (1 + 2b)r)$, and this implies $ma^n \leq (1 + 2b)^n$.

10. Let $(S, \mathcal{O}_d)$ be a topological metric space and let $f : S \longrightarrow \mathbb{R}$ be the function defined by $f(x) = d(x_0, x)$ for $x \in S$. Prove that $f$ is a continuous function between the topological spaces $(S, \mathcal{O}_d)$ and $(\mathbb{R}, \mathcal{O})$.
11. Define the function $f : \mathbb{R} \longrightarrow (0, 1)$ by
$$f(x) = \frac{1}{1 + e^{-x}}$$
for $x \in \mathbb{R}$. Prove that $f$ is a homeomorphism between the topological spaces $(\mathbb{R}, \mathcal{O}_d)$ and $((0, 1), \mathcal{O}_d \restriction_{(0,1)})$.
Conclude that completeness is not a topological property.
12. Let $X$ and $Y$ be two separated sets in the topological metric space $(S, \mathcal{O}_d)$ (recall that the notion of separated sets was introduced in Exercise 12 of Chapter 6). Prove that there are two disjoint open sets $L_1, L_2$ in $(S, \mathcal{O}_d)$ such that $X \subseteq L_1$, $Y \subseteq L_2$.
**Solution:** By Corollary 11.17, the functions $d_X$ and $d_Y$ are continuous, $\mathbf{K}(X) = d_X^{-1}(0)$, and $\mathbf{K}(Y) = d_Y^{-1}(0)$. Since $X$ and $Y$ are separated, we have $X \cap d_Y^{-1}(0) = Y \cap d_X^{-1}(0) = \emptyset$. The disjoint sets $L_1 = \{s \in S \mid d_X(s) - d_Y(s) < 0\}$ and $L_2 = \{s \in S \mid d_X(s) - d_Y(s) > 0\}$ are open due to the continuity of $d_X$ and $d_Y$, and $X \subseteq L_1$ and $Y \subseteq L_2$.
13. This is a variant of the $T_4$ separation property of topological metric spaces formulated for arbitrary sets instead of closed sets.
Let $(S, \mathcal{O}_d)$ be a topological metric space and let $U_1$ and $U_2$ be two subsets of $S$ such that $U_1 \cap \mathbf{K}(U_2) = \emptyset$ and $U_2 \cap \mathbf{K}(U_1) = \emptyset$. There exists two open, disjoint subsets $V_1$ and $V_2$ of $S$ such that $U_i \subseteq V_i$ for $i = 1, 2$.
**Solution:** Define the disjoint open sets
$$V_1 = \{x \in S \mid d(x, U_1) < d(x, U_2)\},$$
$$V_2 = \{x \in S \mid d(x, U_2) < d(x, U_1)\}.$$
We have $U_1 \subseteq V_1$ because, for $x \in U_1$, $d(x, U_1) = 0$ and $d(x, U_2) > 0$ since $x \notin \mathbf{K}(U_2)$. Similarly, $U_2 \subseteq V_2$.
14. Prove that, for every sequence $\mathbf{x}$ of real numbers, we have $\liminf \mathbf{x} = -\limsup(-\mathbf{x})$.
15. Find $\liminf \mathbf{x}$ and $\limsup \mathbf{x}$ for $\mathbf{x} = (x_0, \ldots, x_n, \ldots)$, where
a) $x_n = (-1)^n \cdot n$,
b) $x_n = \frac{(-1)^n}{n}$.
16. Let $(S, \mathcal{O}_d)$ be a topological metric space. Prove that $\mathbf{x} = (x_0, x_1, \ldots)$ is a Cauchy sequence if and only if for every $\epsilon > 0$ there exists $n \in \mathbb{N}$ such that $d(x_n, x_{n+m}) < \epsilon$ for every $m \in \mathbb{N}$.
17. Let $(S, \mathcal{O}_d)$ be a topological metric space. Prove that if every bounded subset of $S$ is compact, then $(S, \mathcal{O}_d)$ is complete.
18. Prove that if $(S_1, d_1), \ldots, (S_n, d_n)$ are complete metric spaces, then their product is a complete metric space.
19. Prove that a topological metric space $(S, \mathcal{O}_d)$ is complete if and only if for every nonincreasing sequence of nonempty closed sets $\mathbf{S} = (S_0, S_1, \ldots)$ such that $\lim_{n\to\infty} diam(S_n) = 0$, we have $\bigcap_{i\in\mathbb{N}} S_i \neq \emptyset$.

20. Prove that $\mathbf{x} = (x_0, x_1, \ldots)$ is a Cauchy sequence in a topological ultrametric space $(S, \mathcal{O}_d)$ if and only if $\lim_{n\to\infty} d(x_n, x_{n+1}) = 0$.
21. Let $(S, d)$ and $(T, d')$ be two metric spaces. Prove that every similarity $f : S \longrightarrow T$ is a homeomorphism between the metric topological spaces $(S, \mathcal{O}_d)$ and $(T, \mathcal{O}_{d'})$.
22. Let $(S, \mathcal{O}_d)$ be a complete topological metric space and let $f : B(x_0, r) \longrightarrow S$ be a contraction such that $d(f(x), f(y)) \leq kd(x, y)$ for $x, y \in B(x_0, r)$ and $k \in (0, 1)$.
    a) Prove that if $d(f(x_0), x_0)$ is sufficiently small, then the sequence $\mathbf{x} = (x_0, x_1, \ldots)$, where $x_{i+1} = f(x_i)$ for $i \in \mathbb{N}$, consists of points located in $B(x_0, r)$.
    b) Prove that $y = \lim_{n\to\infty} x_n$ exists and $f(y) = y$.
23. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a self-similarity of the topological metric $(\mathbb{R}^n, d_2)$ having similarity ratio $r$. Prove that if $H$ is a Lebesgue-measurable set, then $f(H)$ is also Lebesgue-measurable and $\mu(f(H)) = r^n\mu(H)$, where $\mu$ is the Lebesgue outer measure.

    **Solution:** Suppose initially that $r > 0$. Since $f$ is a homeomorphism, the image on an $n$-dimensional interval $I = \prod_{i=1}^n (a_i, b_i)$ is the $n$-dimensional interval $f(I) = \prod_{i=1}^n (f(a_i), f(b_i))$. The definition of $f$ implies $vol(f(I)) = r^n vol(I)$.

    Since

$$\mu(f(H)) = \inf\left\{\sum vol(f(I)) \mid I \in \mathcal{C}, H \subseteq \bigcup I\right\},$$

    it follows that $\mu(f(H)) \leq r^n\mu(H)$. Note that the inverse of $f$ is a self similarity with ratio $\frac{1}{r}$, which implies $\mu(f(H)) \geq r^n\mu(H)$. Thus, $\mu(f(H)) = r^n\mu(H)$.
24. If $U$ is a subset of $\mathbb{R}^n$ and $\mu$ is the Lebesgue outer measure, then $\mu(U) = \inf\{\mu(L) \mid U \subseteq L, L \text{ is open }\}$.

    **Solution:** The monotonicity of $\mu$ implies $\mu(U) \leq \inf\{\mu(L) \mid U \subseteq L, L \text{ is open }\}$. If $\mu(U) = \infty$, the reverse inequality is obvious. Suppose therefore that $\mu(U) < \infty$. By Equality (6.11) of Chapter 6, we have

$$\mu(U) = \inf\left\{\sum vol(I_j) \mid j \in J, U \subseteq \bigcup_{j\in J} I_j\right\},$$

    so there exists a collection of $n$-dimensional open intervals $\{I_j \mid j \in \mathbb{N}\}$ such that $\sum_{j\in\mathbb{N}} \mu(I_j) < \mu(U) + \epsilon$. Thus, $\mu(U) = \inf\{\mu(L) \mid U \subseteq L, L \text{ is open}\}$.
25. A subset $U$ of $\mathbb{R}^n$ is Lebesgue measurable if and only if for every $\epsilon > 0$ there exist an open set $L$ and a closed set $H$ such that $H \subseteq U \subseteq L$ and $\mu(L - H) < \epsilon$.

## Bibliographical Comments

A number of topology texts emphasize the study of topological metric spaces. We mention [47] and [46]. The proof of Theorem 11.28 originates in [78]. Supplement 9 is a result of K. Falconer [48].

# 12

# Dimensions of Metric Spaces

## 12.1 Introduction

Recent research results attempt to counteract the dimensionality curse by focusing on local dimensionality of data by using the fractal dimension of data.

Subsets of $\mathbb{R}^n$ may have "intrinsic" dimensions that are much lower than $n$. Consider, for example, two distinct vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and the line $L = \{\mathbf{a} + t\mathbf{b} \mid t \in \mathbb{R}\}$. Intuitively, $L$ has the intrinsic dimensionality 1; however, $L$ is embedded in $\mathbb{R}^n$ and from this point of view is an $n$-dimensional object. In this chapter we examine formalisms that lead to the definition of this intrinsic dimensionality.

Difficulties related to the high number of correlated features that occur when data mining techniques are applied to data of high dimensionality are collectively designated as the *dimensionality curse*. In Section 12.2 we discuss properties of the $\mathbb{R}^n$ spaces related to the dimensionality curse and we show how the reality of highly dimensional spaces contradicts the common intuition that we acquire through our common experience with lower dimensional spaces. Higher dimensional spaces are approached using analogies with lower dimensional spaces.

## 12.2 The Dimensionality Curse

The term "dimensionality curse," invented by Richard Bellman in [8], is used to describe the difficulties of exhaustively searching a space of high dimensionality for an optimum value of a function defined on such a space. These difficulties stem from the fact that the size of the sets that must be searched increases exponentially with the number of dimensions. Moreover, phenomena that are at variance with the common human intuition acquired in two- or three-dimensional spaces become more significant. This section is dedicated to a study of these phenomena.

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_12, © Springer-Verlag London Limited 2008

The dimensionality curse impacts many data mining tasks, including classification and clustering. Thus, it is important to realize the limitations that working with high-dimensional data impose on designing data mining algorithms.

In Section C.4, we show that the volume of a sphere of radius $R$ in $\mathbb{R}^n$ is

$$V_n(R) = \frac{\pi^{\frac{n}{2}} R^n}{\Gamma\left(\frac{n}{2}+1\right)}$$

Let $Q_n(\ell)$ be an $n$-dimensional cube in $\mathbb{R}^n$. The volume of this cube is $\ell^n$. Consider the $n$-dimensional closed sphere of radius $R$ that is centered in the center of the cube $Q_n(2R)$ and is tangent to the opposite faces of this cube. We have:

$$\lim_{n\to\infty} \frac{V_n(R)}{2^n R^n} = \frac{\pi^{\frac{n}{2}}}{2^n \Gamma\left(\frac{n}{2}+1\right)} = 0.$$

In other words, as the dimensionality of the space grows, the fraction of the cube volume that is located inside the sphere decreases and tends to become negligible for very large values of $n$.

It is interesting to compare the volumes of two concentric spheres of radii $R$ and $R(1-\epsilon)$, where $\epsilon \in (0,1)$. The volume located between these spheres relative to the volume of the larger sphere is

$$\frac{V_n(R) - V_n(R(1-\epsilon))}{V_n(R)} = 1 - (1-\epsilon)^n,$$

and we have

$$\lim_{n\to\infty} \frac{V_n(R) - V_n(R(1-\epsilon))}{V_n(R)} = 1.$$

Thus, for large values of $n$, the volume of the sphere of radius $R$ is concentrated mainly near the surface of this sphere.

Let $Q_n(1)$ be a unit side-length $n$-dimensional cube, $Q_n(1) = [0,1]^n$, centered in $\mathbf{c}_n = (0.5, \ldots, 0.5) \in R^n$. The $d_2$-distance between the center of the cube $\mathbf{c}_n$ and any of its vertices is:

$$\sqrt{0.5^2 + \cdots 0.5^2} = 0.5\sqrt{n},$$

and this value tends to infinity with the number of dimensions $n$ despite the fact that the volume of the cube remains equal to 1. On the other hand, the distance from the center of the cube to any of its faces remains equal to 0.5. Thus, the $n$-dimensional cube is exhibits very different properties in different directions; in other words the $n$-dimensional cube is an anisotropic object.

An interesting property of the unit cube $Q_n(1)$ is observed in [82]. Let $P = (p, \ldots, p) \in \mathbb{R}^n$ be a point located on the main diagonal of $Q_n(1)$ and let $K$ be the subcube of $Q_n(1)$ that includes $(0, \ldots, 0)$ and $P$ and has a side of length $p$; similarly, let $K'$ be the subcube of $Q_n(1)$ that includes $P$ and

$(1, \ldots, 1)$ and has side of length $1 - p$. The ratio of the volumes $V$ and $V'$ of the cubes $K$ and $K'$ is

$$r(p) = \left(\frac{p}{1-p}\right)^n.$$

To determine the increase $\delta$ of $p$ needed to double the volume of this ratio, we must find $\delta$ such that $\frac{r(p+\delta)}{r(p} = 2$, that is

$$\frac{p(1-p) + \delta(1-p)}{p(1-p) - \delta p} = \sqrt[n]{2}.$$

Equivalently, we have

$$\delta = \frac{p(1-p)(\sqrt[n]{2} - 1)}{1 - p + p\sqrt[n]{2}}.$$

The first factor $\frac{p(1-p)}{1-p+p\sqrt[n]{2}}$ remains almost constant for large values of $n$. However, the second factor $\sqrt[n]{2} - 1$ tends toward 0, which shows that within large dimensionality smaller and smaller moves of the point $p$ are needed to double the ratio of the volumes of the cubes $K$ and $K'$. This suggests that the division of $Q_n(1)$ into subcubes is very unstable. If data classifications are attempted based on the location of data vectors in subcubes, this shows in turn the instability of such classification schemes.

Another interesting example of the counterintuitive behavior of spaces of high dimensionality is given in [17]. Now let $Q_n(1)$ be the unit cube centered in the point $\mathbf{c}_n \in \mathbb{R}^n$, where $\mathbf{c}_n = (0.5, \ldots, 0.5)$. For $n = 2$ or $n = 3$, it is easy to see that every sphere that intersects the sides of $Q_2(1)$ or all faces of $Q_3(1)$ must contain the center of the cube $\mathbf{c}_n$. We shall see that, for sufficiently high values of $n$ a sphere that intersects all $(n-1)$-dimensional faces of $Q_n(1)$ does not necessarily contain the center of $Q_n(1)$.

Consider the closed sphere $B(\mathbf{q}_n, r)$, whose center is the point $\mathbf{q}_n = (q, \ldots, q)$, where $q \in [0, 1]$. Clearly, we have $\mathbf{q}_n \in Q_n(1)$ and $d_2(\mathbf{c}_n, \mathbf{q}_n) = \sqrt{n(q^2 - q + 0.25)}$.

If the radius $r$ of the sphere $B(\mathbf{q}_n, r)$ is sufficiently large, then $B(\mathbf{q}_n, r)$ intersects all faces of $Q_n$. Indeed, the distance from $\mathbf{q}_n$ to an $(n-1)$-dimensional face is no more than $\max\{q, 1-q\}$, which shows that $r \geq \max\{q, 1-q\}$ ensures the nonemptiness of all these intersections. Thus, the inequalities

$$n\,(q - 0.5)^2 > r^2 > \max\{q^2, (1-q)^2\} \tag{12.1}$$

ensure that $B(\mathbf{q}_n, r)$ intersects every $(n-1)$-dimensional face of $Q_n$, while leaving $\mathbf{c}_n$ outside $B(\mathbf{q}_n, r)$. This is equivalent to requiring

$$n > \frac{\max\{q^2, (1-q)^2\}}{(q - 0.5)^2}.$$

For example, if we choose $q = 0.3$, then $n > \frac{0.7^2}{0.2^2} = 12.25$. Thus, in the case of $R^{13}$, Inequality (12.1) amounts to $0.52 > r^2 > 0.49$. Choosing $r = \frac{\sqrt{2}}{2}$ gives the sphere with the desired "paradoxical" property.

The examples discussed in this section suggest that precautions and sound arguments are needed when trying to extrapolate familiar properties of two- or three-dimensional spaces to spaces of higher dimensionality.

## 12.3 Inductive Dimensions of Topological Metric Spaces

We present two variants of the inductive dimensions of topological metric spaces: the *small inductive dimension* $ind(S, \mathcal{O}_d)$ and the *large inductive dimension* $IND(S, \mathcal{O}_d)$. Informally, these dimensions capture the intuitive idea that a sphere $B(x, r)$ in $\mathbb{R}^{n+1}$ has a border that is $n$-dimensional. They are defined by inductive definitions, which we present next.

**Definition 12.1.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. The* large inductive dimension *of $(S, \mathcal{O}_d)$ is a member of the set $\{n \in \mathbb{Z} \mid n \geq -1\} \cup \{\infty\}$ defined by:*

(i) *If $S = \emptyset$ and $\mathcal{O}_d = \{\emptyset\}$, then $IND(S, \mathcal{O}_d) = -1$.*
(ii) *$IND(S, \mathcal{O}_d) \leq n$ for $n \geq 0$ if, for every closed set $H$ and every open set $L$ such that $H \subseteq L$, there exists an open set $V$ such that $H \subseteq V \subseteq L$ such that $IND(\partial V, \mathcal{O}_d \restriction_{\partial V}) \leq n - 1$.*
(iii) *$IND(S, \mathcal{O}_d) = n$ if $IND(S, \mathcal{O}_d) \leq n$ and $IND(S, \mathcal{O}_d) \not\leq n - 1$.*
(iv) *If there is no integer $n \geq -1$ such that $IND(S, \mathcal{O}_d) = n$, then $IND(S, \mathcal{O}_d) = \infty$.*

An immediate equivalent definition of $IND(S, \mathcal{O}_d)$ is contained by the next theorem.

**Theorem 12.2.** *If $IND(S, \mathcal{O}_d) \in \mathbb{Z}$, then $IND(S, \mathcal{O}_d)$ is the smallest integer $n$ such that $n \geq -1$, and for every closed set $H$ and every open set $L$ of the topological metric space $(S, \mathcal{O}_d)$ such that $H \subseteq L$, there exists an open set $V$ such that $H \subseteq V \subseteq L$ such that $IND(\partial V, \mathcal{O}_d \restriction_{\partial V}) \leq n - 1$.*

*Proof.* The statement is an immediate consequence of Definition 12.1. □

If we relax the requirement of Definition 12.1 by asserting the existence of the set $V$ only when the closed set $H$ is reduced to an element of $S$, we obtain the following definition of the small inductive dimension.

**Definition 12.3.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. The* small inductive dimension *of $(S, \mathcal{O}_d)$ is a member of the set $\{n \in \mathbb{Z} \mid n \geq -1\} \cup \{\infty\}$ defined by:*

(i) *If $S = \emptyset$ and $\mathcal{O}_d = \{\emptyset\}$, then $ind(S, \mathcal{O}_d) = -1$.*
(ii) *$ind(S, \mathcal{O}_d) \leq n$, where $n \geq 0$, if for $x \in S$ and every open set $L$ that contains $x$, there exists an open set $V$ such that $x \in V \subseteq L$ such that $ind(\partial V, \mathcal{O}_d \restriction_{\partial V}) \leq n - 1$.*
(iii) *$ind(S, \mathcal{O}_d) = n$ if $IND(S, \mathcal{O}_d) \leq n$ and $ind(S, \mathcal{O}_d) \not\leq n - 1$.*

(iv) *If there is no integer* $n \geq -1$ *such that* $ind(S, \mathcal{O}_d) = n$, *then* $ind(S, \mathcal{O}_d) = \infty$.

**Theorem 12.4.** *If* $ind(S, \mathcal{O}_d) \in \mathbb{Z}$, *then* $ind(S, \mathcal{O}_d)$ *is the smallest integer* $n$ *such that* $n \geq -1$, *and for every* $x \in S$ *and every open set* $L$ *that contains* $x$, *there is an open set* $V$ *such that* $x \in V \subseteq L$ *and* $ind(\partial V, \mathcal{O}_d \restriction_{\partial V}) \leq n - 1$.

*Proof.* The statement is an immediate consequence of Definition 12.3. □

Since $\{x\}$ is a closed set for every $x \in S$, it is clear that, for every topological metric space $(S, \mathcal{O}_d)$, we have $ind(S, \mathcal{O}_d) \leq IND(S, \mathcal{O}_d)$.

If there is no risk of confusion, we denote $ind(S, \mathcal{O}_d)$ and $IND(S, \mathcal{O}_d)$ by $ind(S)$ and $IND(S)$, respectively.

**Definition 12.5.** *A topological metric space* $(S, \mathcal{O}_d)$ *is* zero-dimensional *if* $ind(S, \mathcal{O}_d) = 0$.

Clearly, if $IND(S, \mathcal{O}_d) = 0$, then $(S, \mathcal{O}_d)$ is zero-dimensional.

**Theorem 12.6.** *Let* $(S, \mathcal{O}_d)$ *be a nonempty topological metric space. The space is zero-dimensional if and only if there exists a basis for* $\mathcal{O}_d$ *that consists of clopen sets.*

*Proof.* Suppose that $ind(S) = 0$. By Definition 12.3, for every $x \in S$ and every open set $L$, there is an open set $V$ such that $x \in V \subseteq L$ and $ind(\partial V) \leq -1$, which implies $ind(\partial V) = -1$ and thus $\partial V = \emptyset$. This shows that $V$ is a clopen set and the collection of all such sets $V$ is the desired basis.

Conversely, if there exists a basis $\mathcal{B}$ for $\mathcal{O}_d$ such that each set in $\mathcal{B}$ is clopen, then for every $x \in S$ and open set $L$ there exists $V \in \mathcal{B}$ such that $\partial V = \emptyset$, $x \in V = \mathbf{K}(V) \subseteq L$. This implies $ind(S) = 0$. □

**Theorem 12.7.** *Let* $(S, \mathcal{O}_d)$ *be a zero-dimensional separable topological metric space. If* $H_1$ *and* $H_2$ *are two disjoint closed subsets of* $S$, *there exists a clopen set* $U$ *such that* $H_1 \subseteq U$ *and* $U \cap H_2 = \emptyset$.

*Proof.* Since $ind(S) = 0$, by Theorem 12.6 there exists a base $\mathcal{B}$ of $(S, \mathcal{O}_d)$ that consists of clopen sets.

Let $x \in S$. If $x \notin H_1$, then $x$ belongs to the open set $S - H_1$, so there exists $U_x \in \mathcal{B}$ such that $x \in U_x \subseteq S - H_1$, which implies $U_x \cap H_1 = \emptyset$.

If $x \notin H_2$, a similar set $U_x$ can be found such that $x \in U_x \cap H_2 = \emptyset$. Since every $x$ is in either of the two previous cases, it follows that $\mathcal{U} = \{U_x \mid x \in S\}$ is an open cover of $S$ and each set $U_x$ is disjoint from $H_1$ or $H_2$.

By Theorem 11.31, the separability of $(S, \mathcal{O}_d)$ implies that $\mathcal{U}$ contains a countable subcover, $\{U_{x_1}, U_{x_2}, \ldots\}$. Let $V_1, V_2, \ldots$ be the sequence of clopen sets defined inductively by $V_1 = U_{x_1}$, and $V_n = U_{x_n} - \bigcup_{i=1}^{n-1} V_i$ for $n \geq 1$. The sets $V_i$ are pairwise disjoint, $\bigcup_{i \geq 0} V_i = S$, and each set $V_i$ is disjoint from $H_1$ or $H_2$.

Let $U = \bigcup\{V_i \mid V_i \cap H_2 = \emptyset\}$. The set $U$ is open, $H_1 \subseteq U$, and $U \cap H_2 = \emptyset$. Note that the set $U$ is also closed because $S - U = \bigcup\{V_i \mid V_i \cap H_2 \neq \emptyset\}$ is also open. This means that $U$ is clopen and satisfies the conditions of the theorem. □

Theorem 12.7 can be restated by saying that in a zero-dimensional space $(S, \mathcal{O}_d)$, for any two disjoint closed subsets of $S$, $H_1$ and $H_2$, there exist two disjoint clopen subsets $U_1$ and $U_2$ such that $H_1 \subseteq U_1$ and $H_2 \subseteq U_2$. This follows from the fact that we can choose $U_1 = U$ and $U_2 = S - U$.

**Corollary 12.8.** *Let $(S, \mathcal{O}_d)$ be a zero-dimensional separable topological metric space. If $H$ is a closed set and $L$ is an open set such that $H \subseteq L$, then there exists a clopen set $U$ such that $H \subseteq U \subseteq L$.*

*Proof.* This statement is immediately equivalent to Theorem 12.7. □

**Corollary 12.9.** *Let $(S, \mathcal{O}_d)$ be a separable topological metric space. We have $ind(S) = 0$ if and only if $IND(S) = 0$.*

*Proof.* We saw that $IND(S) = 0$ implies $ind(S) = 0$. Suppose that $ind(S) = 0$. By Corollary 12.9, if $H$ is a closed set and $L$ is an open set such that $H \subseteq L$, then there exists a clopen set $U$ such that $H \subseteq U \subseteq L$. This implies $IND(S) = 0$ by Theorem 12.2. □

*Example 12.10.* If $(S, \mathcal{O}_d)$ is a nonempty topological ultrametric space, then $ind(S) = 0$ because the collection of open spheres is a basis for $\mathcal{O}_d$ that consists of clopen sets (see Corollary 11.29).

*Example 12.11.* For any nonempty, finite topological metric space $(S, \mathcal{O}_d)$, we have $ind(S) = 0$. Indeed, consider the a open sphere $C(x, \epsilon)$. If we choose $\epsilon < \min\{d(x, y) \mid x, y \in S \text{ and } x \neq y\}$, then each open sphere consists of $\{x\}$ itself and thus is a clopen set.

*Example 12.12.* Let $\mathbb{Q}$ be the set of rational numbers and let $\mathbb{I} = \mathbb{R} - \mathbb{Q}$ be the set of irrational numbers. Consider the topological metric spaces $(\mathbb{Q}, \mathcal{O}')$ and $(\mathbb{I}, \mathcal{O}'')$, where the topologies $\mathcal{O}'$ and $\mathcal{O}''$ are obtained by restricting the usual metric topology $\mathcal{O}_d$ of $\mathbb{R}$ to $\mathbb{Q}$ and $\mathbb{I}$, respectively. We claim that $ind(\mathbb{Q}, \mathcal{O}') = ind(\mathbb{I}, \mathcal{O}'') = 0$.

Let $r$ be a rational number and let $\alpha$ be an irrational positive number. Consider the set $D(r, \alpha) = \{q \in \mathbb{Q} \mid |r - q| < \alpha\}$. It is easy to see that the collection $\{D(r, \alpha) \mid r \in \mathbb{Q}, \alpha \in \mathbb{I}\}$ is a basis for $(\mathbb{Q}, \mathcal{O}')$. We have

$$\begin{aligned}\partial D(r, \alpha) &= \{q \in \mathbb{Q} \mid |q - r| \leq \alpha\} \cap \{q \in \mathbb{Q} \mid |q - r| \geq \alpha\} \\ &= \{q \in \mathbb{Q} \mid |q - r| = \alpha\} = \emptyset\end{aligned}$$

because the difference of two rational numbers is a rational number. Therefore, the sets of the form $D(r, \alpha)$ are clopen and $ind(\mathbb{Q}, \mathcal{O}') = 0$.

Let $r$ and $p$ be two rational numbers. Consider the set of irrational numbers

$$E(r,p) = \{\alpha \in \mathbb{I} \mid |r-\alpha| < p\}.$$

We claim that the collection $\mathcal{E} = \{E(r,p) \mid r,p \in \mathbb{Q}\}$ is a basis for $(\mathbb{I}, \mathcal{O}'')$. Indeed, let $\beta \in \mathbb{I}$, and let $L$ be an open set in $\mathcal{O}''$. There exists an open sphere $C(\beta,u)$ such that $u > 0$ and $C(\beta,u) \subseteq L$. Let $r_1, r_2 \in \mathbb{Q}$ be two rational numbers such that $\beta - u < r_1 < \beta < r_2 < \beta + u$. If we define $r = (r_1 + r_2)/2$ and $p = (r_2 - r_1)/2$, then $\beta \in E(r,p) \subseteq C(\beta,u) \subseteq L$, which proves that $\mathcal{E}$ is indeed a basis. We have

$$\begin{aligned}\partial E(r,p) &= \{\alpha \in \mathbb{I} \mid |r-\alpha| \leq p\} \cap \{\alpha \in \mathbb{I} \mid |r-\alpha| \geq p\} \\ &= \{\alpha \in \mathbb{I} \mid |r-\alpha| = p\} = \emptyset\end{aligned}$$

for reasons similar to the ones given above. The sets in the basis $\mathcal{E}$ are clopen, and therefore $ind(\mathbb{I}, \mathcal{O}'') = 0$.

*Example 12.13.* We have $ind(\mathbb{R}, \mathcal{O}) = 1$. Indeed, by Theorem 11.52, its topological dimension is not 0 and, on the other hand, it has a basis that consists of spheres $C(x,r)$, that are open intervals of the form $(x-r, x+r)$. Clearly, $\partial(x-r, x+r)$ is the finite set $\{-r, r\}$, which has a small inductive dimension equal to 0. Therefore, $ind(\mathbb{R}, \mathcal{O}) = 1$. It is interesting to observe that this shows that the union of two zero-dimensional sets is not necessarily zero-dimensional because $ind(\mathbb{Q}, \mathcal{O}') = ind(\mathbb{I}, \mathcal{O}'') = 0$, as we saw in Example 12.12.

**Theorem 12.14.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $T$ be a subset of $S$. We have $ind(T, \mathcal{O}_d \restriction_T) \leq ind(S, \mathcal{O}_d)$.*

*Proof.* The statement is immediate if $ind(S, \mathcal{O}_d) = \infty$. The argument for the finite case is by strong induction on $n = \dim(S, \mathcal{O}_d) \geq -1$.

For the base case, $n = -1$, the space $(S, \mathcal{O}_d)$ is the empty space $(\emptyset, \{\emptyset\})$, so $T = \emptyset$ and the inequality obviously holds.

Suppose that the statement holds for topological metric spaces of dimension no larger than $n$. Let $(S, \mathcal{O}_d)$ be a metric topological space such that $ind(S, \mathcal{O}_d) = n+1$, $T$ be a subset of $S$, $t$ be an element of $T$, and $L$ be an open set in $(T, \mathcal{O}_d \restriction_T)$ such that $t \in L$.

There is an open set $L_1 \in \mathcal{O}_d$ such that $L = L_1 \cap T$. Since $ind(S, \mathcal{O}_d) = n+1$, $n$ is the least integer such that there is an open set $W \subseteq S$ for which

$$t \in W \subseteq L_1 \tag{12.2}$$

and $ind(\partial W) \leq n$. The set $V = W \cap T$ is an open set in $(T, \mathcal{O}_d \restriction_T)$, and we have

$$t \in V \subseteq L$$

by intersecting the sets involved in Inclusions (12.2) with $T$. Theorem 6.31 implies that

$$\partial_T(V) = \partial_T(W \cap T) \subseteq \partial_S(W)$$

and, by the inductive hypothesis, $ind(\partial_T(V)) \leq ind(\partial_S(W)) \leq n$. Therefore, the small inductive dimension of $T$ cannot be greater than $n+1$, which is the desired conclusion. □

A similar statement involving the large inductive dimension can be shown.

**Theorem 12.15.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $T$ be a subset of $S$. We have $IND(T, \mathcal{O}_d \restriction_T) \leq IND(S, \mathcal{O}_d)$.*

*Proof.* The argument is similar to the one given in the proof of Theorem 12.14. □

We denote $ind(T, \mathcal{O}_d \restriction_U)$ by $ind(T)$.

An extension of Theorem 12.6 is given next.

**Theorem 12.16.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. We have $ind(S) = n$, where $n \geq 0$, if and only if $n$ is the smallest integer such that there exists a basis for $\mathcal{O}_d$ such that for every $B \in \mathcal{B}$ we have $ind(\partial B) \leq n - 1$.*

*Proof.* The necessity of the condition is immediate from Definition 12.3 because the sets $V$ constitute a basis that satisfies the condition.

To prove that the condition is sufficient, note that from the proof of Theorem 11.23 we obtain the existence of two open disjoint sets $V_1$ and $V_2$ such that $\{x\} \subseteq V_1$ and $S - L \subseteq V_2$ because $\{x\}$ and $S - L$ are two disjoint closed sets. This is equivalent to $x \in V_1 \subseteq S - V_2 \subseteq L$ and, because $S - V_2$ is closed, we have $x \in V_1 \subseteq \mathbf{K}(V_1) \subseteq L$. Let $B$ be a set in the basis such that $x \in B \subseteq V_1$. We have $x \in B \subseteq L$ and $ind(\partial B) \leq n - 1$; since $n$ is the least number with this property, we have $ind(S) = n$. □

**Corollary 12.17.** *For every separable topological metric space $(S, \mathcal{O}_d)$, we have $ind(S) = n$, where $n \geq 0$, if and only if $n$ is the smallest integer such that there exists a countable basis for $\mathcal{O}_d$ such that, for every $B \in \mathcal{B}$, we have $ind(\partial B) \leq n - 1$.*

*Proof.* This statement is a consequence of Theorems 12.16 and 6.48. □

The inductive dimensions can be alternatively described using the notion of set separation.

**Definition 12.18.** *Let $(S, \mathcal{O})$ be a topological space, and let $X$ and $Y$ be two disjoint subsets of $S$. The subset $T$ of $S$* separates the sets $X$ and $Y$ *if there exists two open, disjoint sets $L_1$ and $L_2$ in $(S, \mathcal{O})$ such that $X \subseteq L_1$, $Y \subseteq L_2$, and $T = S - (L_1 \cup L_2)$.*

It is clear that if $T$ separates $X$ and $Y$, then $T$ must be a closed subset of $S$.

Observe that the empty set separates the sets $X$ and $Y$ if and only if the space $S$ is the union of two open disjoint sets $L_1$ and $L_2$ such that $X \subseteq L_1$ and $Y \subseteq L_2$. Since $L_1$ is the complement of $L_2$, both $L_1$ and $L_2$ are clopen sets.

**Theorem 12.19.** *Let $(S, \mathcal{O})$ be a topological space, and let $X$ and $Y$ be two disjoint subsets of $S$. The set $T$ separates the sets $X$ and $Y$ if and only if the following conditions are satisfied:*
(i) *$T$ is a closed set in $(S, \mathcal{O})$, and*
(ii) *there exist two disjoint sets $K_1$ and $K_2$ that are open in the subspace $S-T$ such that $X \subseteq K_1$, $Y \subseteq K_2$, and $S - T = K_1 \cup K_2$.*

*Proof.* Suppose that $T$ separates the sets $X$ and $Y$. It is clear that we have both $X \subseteq S - T$ and $Y \subseteq S - T$. We already observed that $T$ is closed, and so $S - T$ is open. Therefore, the sets $L_1$ and $L_2$ considered in Definition 12.18 are open in $S - T$ and the second condition is also satisfied.

Conversely, suppose that conditions (i) and (ii) are satisfied. Since $T$ is closed, $S - T$ is open. Since $K_1$ and $K_2$ are open in $S - T$, they are open in $(S, \mathcal{O})$, so $T$ separates $X$ and $Y$. □

**Theorem 12.20.** *Let $(S, \mathcal{O})$ be a topological space, $H$ be a closed set, and $L$ be an open set such that $H \subseteq L$. The set $T$ separates the disjoint sets $H$ and $S - L$ if and only if there exists an open set $V$ and a closed set $W$ such that the following conditions are satisfied:*
(i) *$H \subseteq V \subseteq W \subseteq S - L$ and*
(ii) *$T = W - V$.*

*Proof.* Suppose that $T$ separates $H$ and $S - L$. There are two disjoint open sets $L_1$ and $L_2$ such that $H \subseteq L_1$, $S - L \subseteq L_2$, and $T = S - (L_1 \cup L_2)$. This implies $S - L_2 \subseteq L$, and $T = (S - L_1) \cap (S - L_2)$. Let $V = L_1$ and $W = S - L_2$. Since $L_1$ and $L_2$ are disjoint, it is clear that $V \subseteq W$. Also, $T = (S - V) \cap W = W - V$.

Conversely, if the conditions of the theorem are satisfied, then $T$ separates $H$ and $S - L$ because $V$ and $S - W$ are the open sets that satisfy the requirements of Definition 12.18. □

Using the notion of set separation, we have the following characterization of topological metric spaces having large or small inductive dimension $n$.

**Theorem 12.21.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $n \in \mathbb{N}$.*
(i) *$IND(S) = n$ if and only if $n$ is the smallest integer such that for every closed subset $H$ and open set $L$ of $S$ such that $H \subseteq L$ there exists a set $W$ with $IND(W) \leq n - 1$ that separates $H$ and $L$.*
(ii) *$ind(S) = n$ if and only if $n$ is the smallest integer such that for any element $x$ of $S$ and any open set $L$ that contains $x$ there exists a set $W$ with $ind(W) \leq n - 1$ that separates $\{x\}$ and $L$.*

*Proof.* Suppose that $IND(S) = n$. By Definition 12.1, $n$ is the smallest integer such that $n \geq -1$, and for every closed set $H$ and every open set $L$ such that $H \subseteq L$, there exists an open set $V$ such that $H \subseteq V \subseteq \mathbf{K}(V) \subseteq L$ such that $IND(\partial V) \leq n - 1$. Let $W = \mathbf{K}(V) - V$. It is clear that $W$ separates $H$ and $L$. Since

$$W = \mathbf{K}(V) - V = \mathbf{K}(V) \cap (S - V) \subseteq \mathbf{K}(V) \cap \mathbf{K}(S - V) = \partial(V),$$

it follows that $IND(W) \leq n - 1$.

Conversely, suppose that $n$ is the least integer such that for any closed set $H$ and open set $L$ such that $H \subseteq L$ there exist an open set $V$ and a closed set $U$ such that $H \subseteq V \subseteq U \subseteq L$, $W = U - V$, and $IND(W) \leq n - 1$. Clearly, we have $\mathbf{K}(V) \subseteq U$ and therefore

$$H \subseteq V \subseteq \mathbf{K}(V) \subseteq L.$$

Note that

$$\begin{aligned} \partial(V) &= \mathbf{K}(V) \cap \mathbf{K}(S - V) \\ &= \mathbf{K}(V) \cap (S - V) \\ &\quad \text{because } S - V \text{ is a closed set} \\ &\subseteq U \cap (S - V) = U - V = W, \end{aligned}$$

which implies $IND(V) \leq n - 1$.

The proof of the second part of the theorem is similar. □

The next statement shows the possibility of lifting the separation of two closed sets from a subspace to the surrounding space.

**Theorem 12.22.** *Let $(S, \mathcal{O}_d)$ be a topological metric space, and let $H_1$ and $H_2$ be two closed and disjoint subsets of $S$. Suppose that $U_1$ and $U_2$ are two open subsets of $S$ such that $H_1 \subseteq U_1$, $H_2 \subseteq U_2$ and $\boldsymbol{K}(U_1) \cap \boldsymbol{K}(U_2) = \emptyset$.*

*If $T \subseteq S$ and the set $K$ separates the sets $T \cap \boldsymbol{K}(U_1)$ and $T \cap \boldsymbol{K}(U_2)$ in the subspace $(T, \mathcal{O}_d \restriction_T)$, then there exists a subset $W$ of $S$, that separates $H_1$ and $H_2$ in $(S, \mathcal{O}_d)$ such that $W \cap T \subseteq K$.*

*Proof.* Since $K$ separates the sets $T \cap \mathbf{K}(U_1)$ and $T \cap \mathbf{K}(U_2)$ in $T$ there are two open, disjoint subsets $V_1$ and $V_2$ of $T$ such that $T \cap \mathbf{K}(U_1) \subseteq V_1$, $T \cap \mathbf{K}(U_2) \subseteq V_2$, and $T - K = V_1 \cup V_2$.

We have

$$\begin{aligned} U_1 \cap V_2 &= U_1 \cap (T \cap V_2) \\ &= (U_1 \cap T) \cap V_2 \\ &\subseteq \mathbf{K}(U_1) \cap T \cap V_2 \\ &\subseteq V_1 \cap V_2 = \emptyset, \end{aligned}$$

and therefore $U_1 \cap \mathbf{K}(V_2) = \emptyset$, because $U_1$ is open (by Theorem 6.8). Therefore, $H_1 \cap \mathbf{K}(V_2) = \emptyset$. Similarly, $H_2 \cap \mathbf{K}(V_1) = \emptyset$. Consequently,

$$\begin{aligned} (H_1 \cup V_1) \cap (\mathbf{K}(H_2 \cup V_2)) &= (H_1 \cup V_1) \cap (\mathbf{K}(H_2) \cup \mathbf{K}(V_2)) \\ &= (H_1 \cup V_1) \cap (H_2 \cup \mathbf{K}(V_2)) = \emptyset \end{aligned}$$

and similarly

$$\mathbf{K}(H_1 \cup V_1) \cap (H_2 \cup V_2) = \emptyset.$$

By Supplement 13 of Chapter 11, there exist two disjoint open sets $Z_1$ and $Z_2$ such that $H_1 \cup V_1 \subseteq Z_1$ and $H_2 \cup V_2 \subseteq Z_2$. Then, the set $W = S - (Z_1 \cup Z_2)$ separates $H_1$ and $H_2$, and $W \cap T \subseteq T - (Z_1 \cup Z_2) \subseteq T - (V_1 \cup V_2) = K$. □

We can now extend Corollary 12.8.

**Corollary 12.23.** *Let $(S, \mathcal{O}_d)$ be a separable topological metric space and let $T$ be a zero-dimensional subspace of $S$. For any disjoint closed sets $H_1$ and $H_2$ in $S$, there exist two disjoint open sets $L_1$ and $L_2$ such that $H_1 \subseteq L_1$, $H_2 \subseteq L_2$, and $T \cap \partial L_1 = T \cap \partial L_2 = \emptyset$.*

*Proof.* By Theorem 11.26, there are two open sets $V_1$ and $V_2$ such that $H_1 \subseteq V_1$, $H_2 \subseteq V_2$, and $\mathbf{K}(V_1) \cap \mathbf{K}(V_2) = \emptyset$. By Theorem 12.7, there exists a clopen subset $U$ of $T$ such that $T \cap \mathbf{K}(V_1) \subseteq U$ and $T \cap \mathbf{K}(V_2) \subseteq T - U$. Therefore, we have

$$\begin{aligned} T - U &\subseteq S - \mathbf{K}(V_1) \subseteq S - V_1 \subseteq S - H_1, \\ U &\subseteq T - \mathbf{K}(V_2) \subseteq S - V_2 \subseteq S - H_2, \end{aligned}$$

which implies that the sets $H_1 \cup U$ and $H_2 \cup (T - U)$ are disjoint.

Let $f, g : S \longrightarrow \mathbb{R}$ be the continuous functions defined by $f(x) = d_{H_1 \cup U}(x)$ and $g(x) = d_{H_2 \cup (T-U)}(x)$ for $x \in S$. The open sets

$$\begin{aligned} L_1 &= \{x \in S \mid f(x) < g(x)\}, \\ L_2 &= \{x \in S \mid f(x) > g(x)\}, \end{aligned}$$

are clearly disjoint. Note that if $x \in H_1$ we have $f(x) = 0$ and $g(x) > 0$, so $H_1 \subseteq L_1$. Similarly, $H_2 \subseteq L_2$.

Since $U$ is closed in $T$, we have $f(x) = 0$ and $g(x) > 0$ for every $x \in U$; similarly, since $T - U$ is closed in $T$, we have $f(x) > 0$ and $g(x) = 0$. Thus, $U \subseteq L_1$ and $T - U \subseteq L_2$, so $T \subseteq L_1 \cup L_2$.

Note that we have the inclusions

$$\begin{aligned} \partial L_1 &= \mathbf{K}(L_1) \cap \mathbf{K}(S - L_1) \\ &\quad \text{(by the definition of the border)} \\ &= \mathbf{K}(L_1) \cap (S - L_1) \\ &\quad \text{(because } S - L_1 \text{ is a closed set)} \\ &\subseteq \mathbf{K}(S - L_2) \cap S - L_1 \\ &\quad \text{(since } L_1 \subseteq S - L_2) \\ &= (S - L_2) \cap (S - L_1) \\ &= S - (L_1 \cup L_2) \\ &\subseteq S - T. \end{aligned}$$

Similarly, we can show that $\partial L_2 \subseteq S - T$, so $T \cap \partial L_1 = T \cap \partial L_2 = \emptyset$. □

**Theorem 12.24.** *Let $T$ be a zero-dimensional, separable subspace of the topological metric space $(S, \mathcal{O}_d)$. If $H_1$ and $H_2$ are disjoint and closed subsets of $S$, then there exists a set $W$ that separates $H_1$ and $H_2$ such that $W \cap T = \emptyset$.*

*Proof.* By Theorem 11.26, there exist two open sets $U_1$ and $U_2$ such that $H_1 \subseteq U_1$, $H_2 \subseteq U_2$, and $\mathbf{K}(U_1) \cap \mathbf{K}(U_2) = \emptyset$.

Since $T$ is zero-dimensional, the empty set separates the sets $T \cap \mathbf{K}(U_1)$ and $T \cap \mathbf{K}(U_1)$ in the space $T$. By Theorem 12.22, there exists a set $W$ of $S$ that separates $H_1$ and $H_2$ in $(S, \mathcal{O}_d)$ such that $W \cap T = \emptyset$, as stipulated in the statement. □

**Theorem 12.25.** *Let $(S, \mathcal{O}_d)$ be a nonempty separable topological metric space that is the union of a countable collection of zero-dimensional closed sets $\{H_n \mid n \in \mathbb{N}\}$. Then, $(S, \mathcal{O}_d)$ is zero-dimensional.*

*Proof.* Let $x \in S$ and let $L$ be an open set such that $x \in L$. By Corollary 11.27, two open sets $U_0$ and $W_0$ exist such that $x \in U_0$, $L \subseteq W_0$, and $\mathbf{K}(U_0) \cap \mathbf{K}(W_0) = \emptyset$.

We define two increasing sequences of open sets $U_0, U_1, \ldots, U_n, \ldots$ and $W_0, W_1, \ldots, W_n, \ldots$ such that

(i) $\mathbf{K}(U_i) \cap \mathbf{K}(W_i) = \emptyset$ for $i \geq 0$;
(ii) $H_i \subseteq U_i \cup W_i$ for $i \geq 1$.

Suppose that we have defined the sets $U_n$ and $W_n$ that satisfy the conditions above. Observe that the disjoint sets $H_{n+1} \cap \mathbf{K}(U_n)$ and $H_{n+1} \cap \mathbf{K}(W_n)$ are closed in the subspace $H_{n+1}$.

Since $\dim(H_{n+1}) = 0$, by Theorem 12.7, there is a clopen set $K$ in $H_{n+1}$ such that $H_{n+1} \cap \mathbf{K}(U_n) \subset K$ and $K \cap (H_{n+1} \cap \mathbf{K}(W_n)) = \emptyset$. Both $K$ and $H_{n+1} - K$ are closed sets in the space $S$ because $H_{n+1}$ is a closed subset of $S$, which implies that the sets $K \cup \mathbf{K}(U_n)$ and $(H_{n+1} - K) \cup \mathbf{K}(W_n)$ are also closed. Moreover, we have

$$(K \cup \mathbf{K}(U_n)) \cap ((H_{n+1} - K) \cup \mathbf{K}(W_n)) = \emptyset,$$

so there exist two open subsets of $S$, $U_{n+1}$ and $W_{n+1}$, such that $K \cup \mathbf{K}(U_n) \subseteq U_{n+1}$, $(H_{n+1} - K) \cup \mathbf{K}(W_n) \subseteq W_{n+1}$, and $\mathbf{K}(U_{n+1}) \cap \mathbf{K}(W_{n+1}) = \emptyset$.

Consider the open sets $U = \bigcup_{n \in \mathbb{N}} U_n$ and $W = \bigcup_{n \in \mathbb{N}} W_n$. It is clear that $U \cap W = \emptyset$ and $U \cup W = S$, so both $U$ and $W$ are clopen. Since $x \in U_0 \subseteq U = S - W \subseteq V$ and $S - V \subseteq W_0 \subseteq W$, it follows that $S$ is zero-dimensional. □

It is interesting to contrast this theorem with Example 12.13, where we observed that $ind(\mathbb{R}, \mathcal{O}) = 1$ and $ind(\mathbb{Q}, \mathcal{O}') = ind(\mathbb{I}, \mathcal{O}'') = 0$. This happens, of course, because the subspaces $\mathbb{Q}$ and $\mathbb{I}$ are not closed.

**Theorem 12.26.** *Let $(S, \mathcal{O}_d)$ be a separable metric space. If $X$ and $Y$ are two subsets of $S$ such that $S = X \cup Y$, $ind(X) \leq n - 1$, and $ind(Y) = 0$, then $ind(S) \leq n$.*

*Proof.* Suppose that $S$ can be written as $S = X \cup Y$ such that $X$ and $Y$ satisfy the conditions of the theorem. Let $x \in S$ and let $L$ be an open set such that $x \in L$. By applying Theorem 12.24 to the closed sets $\{x\}$ and $S - L$, we obtain the existence of a set $W$ that separates $\{x\}$ and $S - L$ such that $W \cap Y = \emptyset$, which implies $W \subseteq X$. Thus, $ind(W) \leq n - 1$, and this yields $ind(S) \leq n$. □

**Theorem 12.27 (The Sum Theorem).** *Let $(S, \mathcal{O}_d)$ be a separable topological metric space that is a countable union of closed subspaces, $S = \bigcup_{i \geq 1}$, where $ind(H_i) \leq n$. Then, $ind(S) \leq n$, where $n \geq 0$.*

*Proof.* The argument is by strong induction on $n$. The base case, $n = 0$, was discussed in Theorem 12.25.

Suppose that the statement holds for numbers less than $n$, and let $S$ be a countable union of closed subspaces of small inductive dimension less than $n$. By Corollary 12.17, each subspace $H_i$ has a countable basis $\mathcal{B}_i$ such that $ind(\partial_{H_i} B_i) \leq n - 1$ for every $B_i \in \mathcal{B}_i$.

Each border $\partial_{H_i} B_i$ is closed in $H_i$ and therefore is closed in $S$ because each $H_i$ is closed. Define $X = \bigcup_{i \geq 1} \bigcup \{\partial B_i \mid B_i \in \mathcal{B}_i\}$. By the inductive hypothesis, we have $ind(X) \leq n - 1$.

Define the sets $K_i = H_i - X$ for $i \geq 1$. The collection $\mathcal{C}_i = \{K_i \cap B \mid B \in \mathcal{B}\}$ consists of sets that are clopen in $K_i$, and therefore, for each nonempty set $K_i$, we have $ind(K_i) = 0$. Let $Y = S - X$. Since $Y$ is a countable union of the closed subspaces $K_i = H_i \cap Y$, it follows that $ind(Y) = 0$. By Theorem 12.26, it follows that $ind(S) \leq n$. □

The next statement complements Theorem 12.26.

**Corollary 12.28.** *Let $(S, \mathcal{O}_d)$ be a separable metric space. If $ind(S) \leq n$, then there exist two subsets $X$ and $Y$ of $S$ such that $S = X \cup Y$, $ind(X) \leq n - 1$, and $ind(Y) = 0$.*

*Proof.* Let $S$ be such that $ind(S) \leq n$ and let $\mathcal{B}$ be a countable basis such that $ind(\partial B) \leq n - 1$ for every $B \in \mathcal{B}$. The existence of such a basis is guaranteed by Corollary 12.17. Define $X = \bigcup \{\partial B \mid B \in \mathcal{B}\}$. By the Sum Theorem, we have $ind(X) \leq n - 1$. If $Y = S - X$, then $\{Y \cap B \mid B \in B\}$ is a base of $Y$ that consists of clopen sets (in $Y$), so $ind(Y) \leq 0$. □

**Theorem 12.29 (The Decomposition Theorem).** *Let $(S, \mathcal{O}_d)$ be a separable metric space such that $S \neq \emptyset$. We have $ind(S) = n$, where $n \geq 0$ if and only if $S$ is the union of $n + 1$ zero-dimensional subspaces.*

*Proof.* This statement follows from Theorem 12.26. □

**Theorem 12.30.** *[The Separation Theorem] Let $(S, \mathcal{O}_d)$ be a separable topological metric space such that $ind(S) \leq n$, where $n \geq 0$. If $H_1$ and $H_2$ are two disjoint closed subsets, then there exist two disjoint open subsets $L_1$ and $L_2$ that satisfy the following conditions:*

(i) $H_1 \subseteq L_1$ *and* $H_2 \subseteq L_2$*;*
(ii) $ind(\partial L_1) \leq n-1$ *and* $ind(\partial L_2) \leq n-1$.

*Proof.* By Theorem 12.26, there exist two subsets $X$ and $Y$ of $S$ such that $S = X \cup Y$, $ind(X) \leq n-1$, and $ind(Y) = 0$. By Corollary 12.23, there exist two disjoint open sets $L_1$ and $L_2$ such that $H_1 \subseteq L_1$, $H_2 \subseteq L_2$, and $Y \cap \partial L_1 = Y \cap \partial L_2 = \emptyset$. Therefore, $\partial L_1 \subseteq X$ and $\partial L_2 \subseteq X$, so $ind(\partial L_1) \leq n-1$ and $ind(\partial L_2) \leq n-1$. □

The next statement is an extension of Theorem 12.24.

**Theorem 12.31.** *Let $T$ be a subspace of a separable topological metric space $(S, \mathcal{O}_d)$ such that $ind(T) = k$, where $k \geq 0$. If $H_1$ and $H_2$ are disjoint closed subsets of $S$, then there exists a subset $U$ of $S$ that separates $H_1$ and $H_2$ such that $ind(T \cap U) \leq k-1$.*

*Proof.* The case $k = 0$ was discussed in Theorem 12.24.

If $k \geq 1$, then $T = X \cup Y$, where $ind(X) = k-1$ and $ind(Y) = 0$. By Theorem 12.24, the closed sets $H_1$ and $H_2$ are separated by a set $W$ such that $W \cap Y = \emptyset$, which implies $W \cap T \subseteq X$. Thus, $ind(W \cap T) \leq k-1$. □

**Theorem 12.32.** *Let $(S, \mathcal{O}_d)$ be a separable topological metric space. We have $ind(S) = IND(S)$.*

*Proof.* We observed already that $ind(S) \leq IND(S)$ for every topological metric space. Thus, we need to prove only the reverse equality, $IND(S) \leq ind(S)$. This clearly holds if $ind(S) = \infty$.

The remaining argument is by induction on $n = ind(S)$. The base case, $n = 0$, was discussed in Corollary 12.9.

Suppose that the inequality holds for numbers less than $n$. If $H_1$ and $H_2$ are two disjoint and closed sets in $S$, then they can be separated by a subset $U$ of $S$ such that $ind(U) \leq n-1$ by Theorem 12.31. By the induction hypothesis, $IND(U) \leq n-1$, so $IND(S) \leq n$. □

## 12.4 The Covering Dimension

**Definition 12.33.** *Let $\mathcal{E}$ be a family of subsets of a set $S$. The* order *of $\mathcal{E}$ is the least number $n$ such that any $n+2$ sets of $\mathcal{E}$ have an empty intersection.*

*The order of $\mathcal{E}$ is denoted by $ord(\mathcal{E})$.*

If $ord(\mathcal{E}) = n$, then there exist $n+1$ sets in $\mathcal{E}$ that have a nonempty intersection. Also, we have $ord(\mathcal{E}) \leq |\mathcal{E}| + 1$.

*Example 12.34.* If $ord(\mathcal{E}) = -1$, this means that any set of $\mathcal{E}$ is empty, so $\mathcal{E} = \{\emptyset\}$.

The order of any partition is 0.

**Definition 12.35.** *A topological metric space $(S, \mathcal{O}_d)$ has the* covering dimension *$n$ if $n$ is the least number such that $n \geq -1$ and every open cover $\mathcal{C}$ of $S$ has a refinement $\mathcal{D}$ that consists of open sets with $ord(\mathcal{D}) = n$. If no such number $n$ exists, then the covering dimension is $\infty$.*

*The covering dimension of $(S, \mathcal{O}_d)$ is denoted by $cov(S, \mathcal{O}_d)$, or just by $cov(S)$, when there is no risk of confusion.*

**Theorem 12.36.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. The following statements are equivalent:*

(i) *$cov(S) \leq n$.*

(ii) *For every open cover $\mathcal{L} = \{L_1, \ldots, L_p\}$ of $(S, \mathcal{O}_d)$, there is an open cover $\mathcal{K} = \{K_1, \ldots, K_p\}$ such that $ord(\mathcal{K}) \leq n$ and $K_i \subseteq L_i$ for $1 \leq i \leq p$.*

(iii) *For every open cover $\mathcal{L} = \{L_1, \ldots, L_{n+2}\}$ there exists an open cover $\mathcal{K} = \{K_1, \ldots, K_{n+2}\}$ such that $\bigcap \mathcal{K} = \emptyset$ and $K_i \subseteq L_i$ for $1 \leq i \leq n+2$.*

(iv) *For every open cover $\mathcal{L} = \{L_1, \ldots, L_{n+2}\}$ there exists a closed cover $\mathcal{H} = \{H_1, \ldots, H_{n+2}\}$ such that $\bigcap \mathcal{H} = \emptyset$ and $H_i \subseteq L_i$ for $1 \leq i \leq n+2$.*

*Proof.* (i) implies (ii): If $cov(S) \leq n$, then for the open cover $\mathcal{L} = \{L_1, \ldots, L_p\}$ of $(S, \mathcal{O}_d)$ there exists an open cover $\mathcal{U}$ that is a refinement of $\mathcal{L}$ such that $ord(\mathcal{U}) \leq n$. We need to derive from $\mathcal{U}$ another open cover that is also a refinement of $\mathcal{L}$, contains the same number of sets as $\mathcal{L}$, and satisfies the other conditions of (ii).

For $U \in \mathcal{U}$, let $i_U$ be the least number $i$ such that $U \subseteq L_i$. Define the open set $K_i = \bigcup\{U \in \mathcal{U} \mid i_U = i\}$ for $1 \leq i \leq p$. Observe that $\mathcal{K} = \{K_1, \ldots, K_p\}$ is an open cover.

An arbitrary element $x \in S$ belongs to at most $n+1$ members of the collection $\mathcal{U}$ because $ord(\mathcal{U}) \leq n$. Observe that $x \in K_i$ only if $x \in U$ for some $U \in \mathcal{U}$, which implies that $x$ belongs to at most $n+1$ members of $\mathcal{K}$. Thus, $ord(\mathcal{K}) \leq n$.

(ii) implies (iii): This implication is immediate.

(iii) implies (iv): Suppose that (iii) holds, so for every open cover $\mathcal{L} = \{L_1, \ldots, L_{n+2}\}$ there exists an open cover $\mathcal{K} = \{K_1, \ldots, K_{n+2}\}$ such that $\bigcap \mathcal{K} = \emptyset$ and $K_i \subseteq L_i$ for $1 \leq i \leq n+2$. Starting from the open cover $\mathcal{K}$, by Supplement 35(b) of Chapter 6, we obtain the existence of the closed cover $\mathcal{H} = \{H_1, \ldots, H_{n+2}\}$ such that $H_i \subseteq K_i$ for $1 \leq n+2$. This implies immediately that $\mathcal{H}$ satisfies the requirements.

(iv) implies (iii): Suppose that (iv) holds, so for every open cover $\mathcal{L} = \{L_1, \ldots, L_{n+2}\}$ there exists a closed cover $\mathcal{H} = \{H_1, \ldots, H_{n+2}\}$ such that $\bigcap \mathcal{H} = \emptyset$ and $H_i \subseteq L_i$ for $1 \leq i \leq n+2$. By Part (b) of Supplement 36 of Chapter 6, there exists an open cover $\mathcal{K} = \{K_1, \ldots, K_{n+2}\}$ such that $K_i \subseteq L_i$ for $1 \leq i \leq n+2$ and $\bigcap \mathcal{K} = \emptyset$.

(iii) implies (ii): Suppose that $(S, \mathcal{O})$ satisfies condition (iii), and let $\mathcal{L} = \{L_1, \ldots, L_p\}$ be an open cover of $(S, \mathcal{O}_d)$. If $p \leq n+1$, then the desired collection is $\mathcal{L}$ itself. Thus, we may assume that $p \geq n+2$.

We need to prove that there exists an open cover $\mathcal{K} = \{K_1, \ldots, K_p\}$ such that $ord(\mathcal{K}) \leq n$ and $K_i \subseteq L_i$ for $1 \leq i \leq p$. This means that we have to show that the intersection on any $n+2$ sets of $\mathcal{K}$ is empty. Without loss of generality, we can prove that the intersection of the first $n+2$ sets of $\mathcal{K}$ is empty.

Consider the open cover $\{L_1, \ldots, L_{n+1}, L_{n+2} \cup \cdots \cup L_p\}$. By (iii), there exists an open cover $\mathcal{Q} = \{Q_1, \ldots, Q_{n+2}\}$ such that $\bigcap \mathcal{Q} = \emptyset$ and $Q_i \subseteq L_i$ for $1 \leq i \leq n+1$ and $Q_{n+2} \subseteq L_{n+2} \cup \cdots \cup L_p$. For $1 \leq i \leq p$, define the open sets

$$K_i = \begin{cases} Q_i & \text{if } 1 \leq i \leq n+1, \\ Q_{n+2} \cap L_i & \text{if } n+2 \leq i \leq p. \end{cases}$$

The collection $\mathcal{K} = \{K_1, \ldots, K_p\}$ is clearly an open cover, $K_i \subseteq L_i$ for $1 \leq i \leq p$, and $\bigcap_{i=1}^{n+2} K_i = \emptyset$.

(ii) implies (i): This implication is immediate. □

**Corollary 12.37.** *Let $(S, \mathcal{O}_d)$ be a nonempty topological metric space. The following statements are equivalent:*

(i) *$cov(S) = 0$.*

(ii) *For all open sets $L_1$ and $L_2$ such that $L_1 \cup L_2 = S$, there exist two disjoint open sets $K_1$ and $K_2$ such that $K_i \subseteq L_i$ for $i \in \{1, 2\}$.*

(iii) *For all open sets $L_1$ and $L_2$ such that $L_1 \cup L_2 = S$ there exist two disjoint closed sets $H_1$ and $H_2$ such that $H_i \subseteq L_i$ for $i \in \{1, 2\}$.*

*Proof.* This corollary is a special case of Theorem 12.36. □

**Theorem 12.38.** *Let $(S, \mathcal{O}_d)$ be a topological metric space. We have $cov(S) = 0$ if and only if $IND(S) = 0$.*

*Proof.* Suppose that $cov(S) = 0$. Let $H_1$ and $H_2$ be two disjoint closed sets. Then $\{S - H_1, S - H_2\}$ is an open cover of $S$. By Part (ii) of Corollary 12.37, there exist two disjoint open sets $K_1$ and $K_2$ such that $K_1 \subseteq S - H_1$ and $K_2 \subseteq S - H_2$. Thus, $K_1 \cup K_2 \subseteq (S - H_1) \cup (S - H_2) = S - (H_1 \cap H_2) = S$, which means that both $K_1$ and $K_2$ are clopen. This implies $IND(S) = 0$.

Conversely, suppose that $IND(S) = 0$, so $ind(S) = 0$. Let $L_1$ and $L_2$ be two open sets such that $L_1 \cup L_2 = S$. The closed sets $S - L_1$ and $S - L_2$ are disjoint, so by Theorem 12.7 there exists a clopen set $U$ such that $S - L_1 \subseteq U$ (that is, $S - U \subseteq L_1$) and $U \cap (S - L_2) = \emptyset$ (that is, $U \subseteq L_2$). Since the sets $S - U$ and $U$ are also closed, it follows that $cov(S) = 0$ by the last part of Corollary 12.37. □

**Theorem 12.39.** *If $(S, \mathcal{O}_d)$ is a separable topological space, then $cov(S) \leq ind(S)$.*

*Proof.* The statement clearly holds if $ind(S) = \infty$. Suppose now that $ind(S) = n$.

By the Decomposition Theorem (Theorem 12.29), $S$ is the union of $n+1$ zero-dimensional spaces, $S = \bigcup_{i=1}^{n+1} T_i$. If $\mathcal{L} = \{L_1, \ldots, L_m\}$ is a finite open cover of $S$, then $\mathcal{C} = \{L_1 \cap T_i, \ldots, L_m \cap T_i\}$ is a finite open cover of the subspace $T_i$. Since $ind(T_i) = 0$, we have $cov(T_i) = 0$ by Theorem 12.38. Therefore, the open cover $\mathcal{C}$ has a finite refinement that consists of disjoint open sets of the form $K_{ij}$ such that $K_{ij} \subseteq L_j$ and $\bigcup_{j=1} K_{ij} \subseteq T_i$. Consequently, the collection $\mathcal{K} = \{B_{ij} \mid 1 \leq i \leq n+1, 1 \leq j \leq m\}$ is a cover of $S$ that refines the collection $\mathcal{L}$. Every subcollection $\mathcal{K}'$ of $\mathcal{K}$ that contains $n+2$ sets must contain two sets that have the same second index, so any such intersection is empty. This allows us to conclude that $cov(S) \leq n = ind(S)$. □

## 12.5 The Cantor Set

We introduce a special subset of the set of real numbers that plays a central role in the dimension theory of metric spaces.

Let $v_n : \{0,1\}^n \longrightarrow \mathbb{N}$ be the function defined by

$$v_n(b_0, b_1, \ldots, b_{n-1}) = 2^{n-1}b_0 + \cdots + 2b_{n-2} + b_{n-1}$$

for every sequence $(b_0, \ldots, b_n) \in \{0,1\}^n$. Clearly, $v_n(b_0, \ldots, b_{n-2}, b_{n-1})$ yields the number designated by the binary sequence $(b_0, \ldots, b_{n-2}, b_{n-1})$. For example, $v_3(110) = 2^2 \cdot 1 + 2^1 \cdot 1 + 0 = 6$.

Similarly, let $w_n : \{0,1,2\}^n \longrightarrow \mathbb{N}$ be the function defined by

$$w_n(b_0, b_1, \ldots, b_{n-1}) = 3^{n-1}b_0 + \cdots + 3b_{n-2} + b_{n-1}$$

for every sequence $(b_0, \ldots, b_n) \in \{0,1,2\}^n$. Then, $w_n(b_0, \ldots, b_{n-2}, b_{n-1})$ is the number designated by the ternary sequence $(b_0, \ldots, b_{n-2}, b_{n-1})$. For example, $w_3(110) = 3^2 \cdot 1 + 3^1 \cdot 1 + 0 = 12$.

Consider a sequence of subsets of $\mathbb{R}$, $E^0, E^1, \ldots$, where $E^0 = [0,1]$ and $E^1$ is obtained from $E^0$ by removing the middle third $(1/3, 2/3)$ of $E^0$. If the remaining closed intervals are $E_0^1$ and $E_1^1$, then $E^1$ is defined by $E^1 = E_0^1 \cup E_1^1$.

By removing the middle intervals from the sets $E_0^1$ and $E_1^1$, four new closed intervals $E_{00}^2, E_{01}^2, E_{10}^2, E_{11}^2$ are created. Let $E^2 = E_{00}^2 \cup E_{01}^2 \cup E_{10}^2 \cup E_{11}^2$.

$E^n$ is constructed from $E^{n-1}$ by removing $2^{n-1}$ disjoint middle third intervals from $E^{n-1}$ (see Figure 12.1). Namely, if $E_{i_0 \cdots i_{n-1}}^n$ is an interval of the set $E^n$, by removing the middle third of this interval, we generate two closed intervals $E_{i_0 \cdots i_{n-1} 0}^{n+1}$ and $E_{i_0 \cdots i_{n-1} 1}^{n+1}$.

In general, $E_n$ is the union of $2^n$ closed intervals

$$E^n = \bigcup_{i_0, \ldots, i_{n-1}} \{E_{i_0, \ldots, i_{n-1}}^n \mid (i_0, \ldots, i_{n-1}) \in \{0,1\}^n\},$$

for $n \geq 0$.

An argument by induction on $n \in \mathbb{N}$ shows that

$$E^n_{i_0\cdots i_{n-1}} = \left[\frac{2w_n(i_0,\ldots,i_{n-1})}{3^n}, \frac{2w_n(i_0,\ldots,i_{n-1})+1}{3^n}\right].$$

Indeed, the equality above holds for $n = 0$. Suppose that it holds for $n$, and denote by $a$ and $b$ the endpoints of the interval $E^n_{i_0\cdots i_{n-1}}$; that is,

$$\begin{aligned} a &= \frac{2w_n(i_0,\ldots,i_{n-1})}{3^n}, \\ b &= \frac{2w_n(i_0,\ldots,i_{n-1})+1}{3^n}. \end{aligned}$$

By the inductive hypothesis, the points that divide $E^n_{i_0\cdots i_{n-1}}$ are

$$\begin{aligned} \frac{2a+b}{3} &= \frac{6w_n(i_0,\ldots,i_{n-1})+1}{3^{n+1}} \\ &= \frac{2w_{n+1}(i_0,\ldots,i_{n-1},0)+1}{3^{n+1}} \end{aligned}$$

and

$$\begin{aligned} \frac{a+2b}{3} &= \frac{6w_n(i_0,\ldots,i_{n-1})+2}{3^{n+1}} \\ &= \frac{2w_{n+1}(i_0,\ldots,i_{n-1},1)}{3^{n+1}}. \end{aligned}$$

Thus, the remaining left third of $E^n_{i_0\cdots i_{n-1}}$ is

$$\begin{aligned} E^{n+1}_{i_0\cdots i_{n-1}0} &= \left[\frac{2w_n(i_0,\ldots,i_{n-1})}{3^n}, \frac{2w_{n+1}(i_0,\ldots,i_{n-1},0)+1}{3^{n+1}}\right] \\ &= \left[\frac{2w_{n+1}(i_0,\ldots,i_{n-1},0)}{3^{n+1}}, \frac{2w_{n+1}(i_0,\ldots,i_{n-1},0)+1}{3^{n+1}}\right], \end{aligned}$$

while the remaining right third is

$$\begin{aligned} E^{n+1}_{i_0\cdots i_{n-1}1} &= \left[\frac{2w_{n+1}(i_0,\ldots,i_{n-1},1)}{3^{n+1}}, \frac{2w_n(i_0,\ldots,i_{n-1})+1}{3^n}\right] \\ &= \left[\frac{2w_{n+1}(i_0,\ldots,i_{n-1},1)}{3^{n+1}}, \frac{2w_{n+1}(i_0,\ldots,i_{n-1},1)+1}{3^{n+1}}\right], \end{aligned}$$

which concludes the inductive argument.

Each number $x$ located in the leftmost third $E^1_0 = [0, 1/3]$ of the set $E_0 = [0, 1]$ can be expressed in base 3 as a number of the form $x = 0.0d_2d_3\cdots$; the number $1/3$, the right extreme of this interval, can be written either as $x = 0.1$ or $x = 0.022\cdots$. We adopt the second representation which allows us to say that all numbers in the rightmost third $E^1_1 = [2/3, 1]$ of $E^0$ have the form $0.2d_2d_3\cdots$ in the base 3.

The argument applies again to the intervals $E^2_{00}, E^2_{01}, E^2_{10}, E^2_{11}$ obtained from the set $E^1$. Every number $x$ in the interval $E^2_{ij}$ can be written in base 3 as $x = 0.i'j'\cdots$, where $i' = 2i$ and $j' = 2j$.

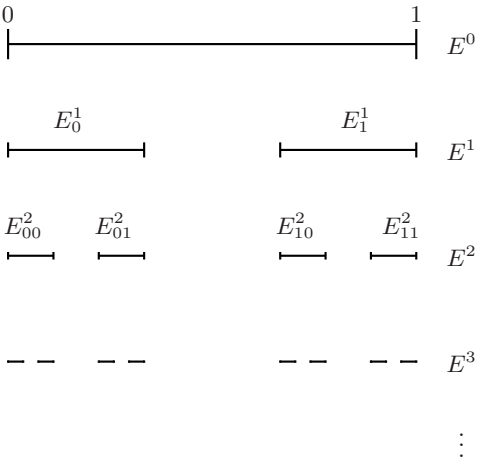



**Fig. 12.1.** Construction of the Cantor dust.

The Cantor set is the intersection

$$C = \bigcap \{E^n \mid n \geq 0\}.$$

Let us evaluate the total length of the intervals of which a set of the form $E_n$ consists. There are $2^n$ intervals of the form $E^n_{i_0 \cdots i_{n-1}}$, and the length of each of these intervals is $\frac{1}{3^n}$. Therefore, the length of $E_n$ is $(2/3)^n$, so this length tends toward 0 when $n$ tends towards infinity. In this sense, the Cantor set is very sparse. Yet, surprisingly, the Cantor set is equinumerous with the interval $[0, 1]$. To prove this fact, observe that the Cantor set consists of the numbers $x$ that can be expressed as

$$x = \sum_{n=1}^{\infty} \frac{a_n}{3^n},$$

where $a_n \in \{0, 2\}$ for $n \geq 1$. For example, $1/4$ is a member of this set since $1/4$ can be expressed in base 3 as $0.020202 \cdots$. Define the function $g : C \longrightarrow [0, 1]$ by $g(x) = y$ if $x = 0.a_1 a_2 \cdots$ (in base 3), where $a_i \in \{0, 2\}$ for $i \geq 1$ and $y = 0.b_1 b_2 \cdots$ (in base 2), where $b_i = a_i/2$ for $i \geq 1$. It is easy to see that this is a bijection between $C$ and $[0, 1]$, which shows that these sets are equinumerous.

We now study the behavior of the sets

$$E^n_{i_0 \cdots i_{n-1}} = \left[ \frac{2w_n(i_0, \ldots, i_{n-1})}{3^n}, \frac{2w_n(i_0, \ldots, i_{n-1}) + 1}{3^n} \right]$$

relative to two mappings $f_0, f_1 : [0, 1] \longrightarrow [0, 1]$ defined by

$$f_0(x) = \frac{x}{3} \text{ and } f_1(x) = \frac{x + 2}{3}$$

for $x \in [0,1]$.

Note that

$$\begin{aligned} f_0(E^n_{i_0\cdots i_{n-1}}) &= \left[\frac{2w_n(i_0,\ldots,i_{n-1})}{3^{n+1}}, \frac{2w_n(i_0,\ldots,i_{n-1})+1}{3^{n+1}}\right] \\ &= \left[\frac{2w_{n+1}(0i_0,\ldots,i_{n-1})}{3^{n+1}}, \frac{2w_{n+1}(0i_0,\ldots,i_{n-1})+1}{3^{n+1}}\right] \\ &= E^{n+1}_{0i_0\cdots i_{n-1}}. \end{aligned}$$

Similarly,

$$f_1(E^n_{i_0\cdots i_{n-1}}) = E^{n+1}_{1i_0\cdots i_{n-1}}.$$

Thus, in general, we have $f_i(E^n_{i_0\cdots i_{n-1}}) = E^{n+1}_{ii_0\cdots i_{n-1}}$ for $i \in \{0,1\}$.

This allows us to conclude that $E^{n+1} = f_0(E^n) \cup f^1(E^n)$ for $n \in \mathbb{N}$. Since both $f_0$ and $f_1$ are injective, it follows that

$$\begin{aligned} C &= \bigcap_{n\geq 1} E^n = \bigcap_{n\geq 0} E^{n+1} \\ &= \bigcap_{n\geq 0} [f_0(E^n) \cup f_1(E^n)] \\ &= \left(\bigcap_{n\geq 0} f_0(E^n)\right) \cup \left(\bigcap_{n\geq 0} f_1(E^n)\right) \\ &= f_0\left(\bigcap_{n\geq 0} E^n\right) \cup f_1\left(\bigcap_{n\geq 0} E^n\right). \end{aligned}$$

In Figure 12.2 we show how sets of the form $E^2_{ij}$ are mapped into sets of the form $E^3_{ijk}$ by $f_0$ (represented by plain arrows) and $f_1$ (represented by dashed arrows).

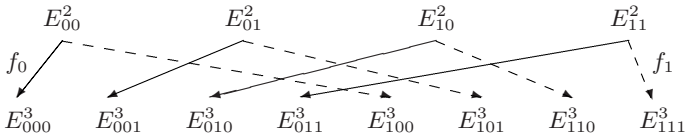



**Fig. 12.2.** Mapping sets $E^2_{ij}$ into sets $E^3_{ijk}$.

**Theorem 12.40.** *The small inductive dimension of the Cantor set is* 0.

*Proof.* We saw that

$$C = \bigcap_{n\in\mathbb{N}} E^n = \bigcap_{n\in\mathbb{N}} \bigcup_{i_0\cdots i_{n-1}} E^n_{i_0\cdots i_{n-1}}.$$

The sets $C\cap E^n_{i_0\cdots i_{n-1}}$ form a base for the open sets of the subspace $C$ of $(\mathbb{R}, \mathcal{O})$. Note that the length of a closed interval $E^n_{i_0\cdots i_{n-1}}$ is $\frac{1}{3^n}$ and the distance between two distinct intervals $E^n_{i_0\cdots i_{n-1}}$ and $E^n_{j_0\cdots j_{n-1}}$ is at least $\frac{1}{3^n}$. Thus, $C \cap E^n_{i_0\cdots i_{n-1}}$ is closed in $C$. On the other hand, the same set is also open because

$$C \cap E^n_{i_0\cdots i_{n-1}} = C \cap \left(a - \frac{1}{3^n}, b + \frac{1}{3^n}\right),$$

where

$$a = \frac{2w_n(i_0, \ldots, i_{n-1})}{3^n},$$
$$b = \frac{2w_n(i_0, \ldots, i_{n-1}) + 1}{3^n}.$$

If $x \in C$, then $C \cap E^n_{i_0\cdots i_{n-1}} \subseteq C \cap S(x, r)$ provided that $\frac{1}{3^n} < r$. This shows that $C$ has a basis that consists of clopen sets, so $ind(C) = 0$ by Theorem 12.6. □

## 12.6 The Box-Counting Dimension

The box-counting dimension reflects the variation of the results of measuring a set at a diminishing scale, which allows the observation of progressively smaller details.

Let $(S, \mathcal{O}_d)$ be a topological metric space and let $T$ be a precompact set. For every positive $r$, there exists an $r$-net for $T$; that is, a finite subset $N_r$ of $S$ such that $T \subseteq \bigcup\{C(x, r) \mid x \in N_r\}$ for every $r > 0$. Denote by $n_T(r)$ the *smallest* size of an $r$-net of $T$. It is clear that $r < r'$ implies $n_T(r) \geq n_T(r')$.

**Definition 12.41.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $T$ be a precompact set.*

*The* upper box-counting dimension *of $T$ is the number*

$$ubd(T) = \limsup_{r\to 0} \frac{n_T(r)}{\log \frac{1}{r}}.$$

*The* lower box-counting dimension *of $T$ is the number*

$$lbd(T) = \liminf_{r\to 0} \frac{n_T(r)}{\log \frac{1}{r}}.$$

*If $ubd(T) = lbd(T)$, we refer to their common values as the* box-counting dimension *of $T$, denoted by $bd(T)$.*

*Example 12.42.* Let $T = \{0\} \cup \{\frac{1}{n} \mid n \geq 1\}$ be a subset of $\mathbb{R}$. The interval $[0, r]$ contains almost all members of $T$ because if $n \geq \lceil \frac{1}{r} \rceil$, we have $\frac{1}{n} \in [0, r]$.

It is easy to verify that

$$\frac{1}{n-1} - \frac{1}{n} > \frac{1}{n} - \frac{1}{n+1},$$

for $n > 1$. Note that $\frac{1}{n-1} - \frac{1}{n} > r$ is equivalent to $n^2 - n - \frac{1}{r} < 0$, and this happens when

$$n < \frac{1 + \sqrt{1 + \frac{4}{r}}}{2}.$$

Thus, each number of the form $\frac{1}{n}$ for

$$n \leq n_0 = \left\lfloor \frac{1 + \sqrt{1 + \frac{4}{r}}}{2} \right\rfloor$$

requires a distinct interval of size $r$ to be covered.

The portion of $T$ that is located to the left of $\frac{1}{n_0}$ and ends with the number $r$ has length $\frac{1}{n_0} - r$ and can be covered with no more than $\frac{1}{rn_0} - 1$ intervals of length $r$. The least number of intervals of length $r$ that are needed has

$$\frac{1}{rn_0} + \left\lfloor \frac{1 + \sqrt{1 + \frac{4}{r}}}{2} \right\rfloor$$

as an upper bound, a number that has the order of magnitude $\Theta(r^{-1/2})$. Thus, $ubd(T) \leq \frac{1}{2}$.

The notion of an $r$-net is related to two other notions, which we introduce next.

**Definition 12.43.** *Let $(S, d)$ be a metric space, $T$ be a subset of $S$, and let $r$ be a positive number.*

*A collection $\mathcal{C}$ of subsets of $S$ is an* $r$-cover *of $T$ of $S$ if, for every $C \in \mathcal{C}$, $diam_d(C) \leq 2r$ and $T \subseteq \bigcup \mathcal{C}$;*

*A subset $W$ of $T$ is* $r$-separated *if, for every $x, y \in W$, $x \neq y$ implies $d(x, y) > r$. The cardinality of the* largest $r$*-separated subset $W$ of $T$ is denoted by $\wp_T(r)$ and will be referred to as the* $r$-separation number *of $T$.*

Observe that an $r$-net for a set $T$ is an $r$-cover.

*Example 12.44.* Consider the metric space $([0, 1]^2, d_2)$, where $d_2$ is the Euclidean metric. Since the area of a circle $B_{d_2}(x, r)$ is $\pi r^2$, it follows that for any $2r$-cover $\mathcal{C}$ that consists of circles, we have $\pi \cdot r^2 \cdot |\mathcal{C}| \geq 1$. Thus, a cover of this type of $([0, 1]^2, d_2)$ must contain at least $\frac{1}{\pi \cdot r^2}$ circles.

In general, the volume of a sphere $B_{d_2}(x, r)$ in $\mathbb{R}^n$ is

$$\frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} r^n,$$

which means that in the metric space $([0,1]^n, d_2)$, a cover by spheres of radius $r$ contains at least

$$\frac{\Gamma\left(\frac{n}{2}+1\right)}{\pi^{\frac{n}{2}} r^n}$$

spheres.

*Example 12.45.* Let $W = \{w_1, \ldots, w_n\}$ be an $r$-separated subset of the interval $[0,1]$, where $w_1 < \cdots < w_n$. We have $1 \geq w_n - w_1 \geq (n-1)r$, so $n \leq \frac{1}{r} + 1$. This implies

$$\wp_{[0,1]}(r) = \left\lfloor \frac{1}{r} + 1 \right\rfloor.$$

*Example 12.46.* Let $T = \{0\} \cup \{\frac{1}{n} \mid n \in \mathbb{N}_1\}$. We seek to determine an upper bound for $\wp_T(r)$. Note that if $p > n$, then

$$\frac{1}{n} - \frac{1}{p} \geq \frac{1}{n} - \frac{1}{n+1}.$$

By the Mean Value Theorem, there exists $c \in (n, n+1)$ such that

$$\frac{1}{n} - \frac{1}{n+1} = \frac{1}{c^2}$$

and therefore

$$\frac{1}{n^2} > \frac{1}{n} - \frac{1}{n+1} > \frac{1}{(n+1)^2}.$$

Let $n_1$ be the largest number such that $\frac{1}{(n_1+1)^2} \geq r$. Then, an $r$-separated subset of $T$ has at least $n_1$ elements; thus, the number $\frac{1}{\sqrt{r}}$ is a lower bound for the number of elements of an $r$-separating set.

**Theorem 12.47.** *Let $T$ be a subset of a metric space $(S, d)$. The following statements are equivalent:*

(i) *For each $r > 0$, there exists an $r$-net for $T$.*
(ii) *For each $r > 0$, every $r$-separated subset of $T$ is finite.*
(iii) *For each $r > 0$, there exists a finite $r$-cover of $T$.*

*Proof.* (i) implies (ii): Let $W$ be an $r$-separated subset of $T$ and let $N_{\frac{r}{2}}$ be an $\frac{r}{2}$-net for $T$. By the definition of $\frac{r}{2}$-nets, for each $w \in W$ there exists $t \in N_{\frac{r}{2}}$ such that $d(w,t) < \frac{r}{2}$; that is, $w \in C\left(t, \frac{r}{2}\right)$. Note that each sphere $C(t, \frac{r}{2})$ contains at most one member of $w$ because $W$ is an $r$-separated subset of $T$. The finiteness of $T$ implies that $W$ is finite too.

(ii) implies (iii): Let $W = \{w_1, \ldots, w_n\}$ be a maximal finite $r$-separated subset of $T$. If $t \in T$, then there exists $w_i \in W$ such that $d(t, w_i) \leq r$ since otherwise the maximality of $W$ would be contradicted. Thus, $T \subseteq \bigcup_{i=1}^{n} B(w_i, r)$, each set $B(w_i, r)$ has a diameter of $2r$, and this implies that $\{B(w_i, r) | 1 \leq i \leq n\}$ is an $r$-cover of $T$.

(iii) implies (i): Let $\mathcal{D} = \{D_1, \ldots, D_n\}$ be a finite $\frac{r+\epsilon}{2}$-cover of $T$, where $\epsilon > 0$. Select $y_i \in D_i$ for $1 \leq i \leq n$, and define the set $Y = \{y_1, \ldots, y_n\}$. Since the diameter of every set $D_i$ is not larger than $r + \epsilon$, for every $t \in T$ there exists $y_i$ such that $d(t, y_i) \leq r + \epsilon$ for every $\epsilon > 0$, so $d(t, y_i) < r$. Therefore, $Y$ is an $r$-net for $T$. □

The connection between the numbers $n_T(r)$ and $\wp_T(r)$ is discussed next.

**Corollary 12.48.** *For every precompact set $T$ of a topological metric space $(S, \mathcal{O}_d)$, we have*

$$n_T(r) \leq \wp_T(r) \leq n_T\left(\frac{r}{2}\right),$$

*for every positive $r$.*

*Proof.* The first inequality follows from the proof of the first implication in Theorem 12.47. The second is a consequence of the last two implications of the same theorem. □

The open spheres of radius $r$ in the definition of a box-counting dimension of a precompact set $T$ can be replaced with arbitrary sets of diameter $2r$. Indeed, suppose $n_T(r)$ is the smallest number of open spheres of radius $r$ that cover $T$ and $n_T(2r)'$ is the least number of sets of diameter $2r$ that cover $T$. Since each sphere of radius $r$ has diameter $2r$, we have $n_T(2r)' \leq n_T(r)$. Observe that each set of diameter $2r$ that intersects $T$ is enclosed in an open sphere with radius $2r$ centered in $T$, so $n_T(2r) \leq n_T(2r)'$. The inequalities $n_T(2r) \leq n_T(2r)' \leq n_T(r)$ imply that the replacement previously mentioned does not affect the value of the box dimension. For example, open spheres can be replaced by closed spheres without affecting the value of the box-counting dimension.

**Theorem 12.49.** *Let $T$ be a subset of a topological metric space $(S, \mathcal{O}_d)$. We have $ubd(\boldsymbol{K}(T)) = ubd(T)$ and $lbd(\boldsymbol{K}(T)) = lbd(T)$.*

*Proof.* Let $\{B(x_1, r), \ldots, B(x_n, r)\}$ be a finite collection of closed spheres such that $T \subseteq \bigcup_{i=1}^{n} B(x_i, r)$. Clearly, we have $\mathbf{K}(T) \subseteq \bigcup_{i=1}^{n} B(x_i, r)$. Thus, a finite collection of closed spheres covers $T$ if and only if it covers $\mathbf{K}(T)$. The conclusion follows immediately. □

## 12.7 The Hausdorff-Besicovitch Dimension

The Hausdorff-Besicovitch measure plays a fundamental role in the study of fractals. The best-known definition of fractals was formulated by B. Mandelbrot [94], who is the founder of this area of mathematics, and states that

a fractal is a geometrical object whose Hausdorff-Besicovitch dimension is greater than its small inductive dimension. The most famous example is the Cantor set whose small inductive dimension is 0 (by Theorem 12.40) and whose Hausdorff-Besicovitch dimension is $\frac{\ln 2}{\ln 3}$, as we shall prove below.

Recall that a collection $\mathcal{C}$ of subsets of a metric space $(S, d)$ is an *r-cover* of a subset $U$ of $S$ if, for every $C \in \mathcal{C}$, $diam_d(C) \leq 2r$ and $U \subseteq \bigcup \mathcal{C}$.

Let $(S, d)$ be a metric space and let $\mathfrak{C}_r(U)$ be the collection of all countable $r$-covers for a set $U$. Observe that $r_1 \leq r_2$ implies $\mathfrak{C}_{r_1}(U) \subseteq \mathfrak{C}_{r_2}(U)$ for $r_1, r_2 \in \mathbb{R}_{>0}$.

Let $s$ be a positive number. We shall use the outer measure $HB_r^s$ obtained by applying Method I (see Theorem 6.127) to the function $f : \mathcal{C} \longrightarrow \mathbb{R}_{\geq 0}$ given by $f(C) = (diam(C))^s$ for $C \in \mathcal{C}$, which is given by

$$HB_r^s(U) = \inf_{\mathcal{C} \in \mathfrak{C}_r(U)} \sum \{(diam(C))^s \mid C \in \mathcal{C}\}.$$

The function $HB_r^s(U)$ is antimonotonic with respect to $r$. Indeed, if $r_1 \leq r_2$, then $\mathfrak{C}_{r_1}(U) \subseteq \mathfrak{C}_{r_2}(U)$, so

$$\inf_{\mathcal{C} \in \mathfrak{C}_{r_2}(U)} \sum \{(diam(C))^s \mid C \in \mathcal{C}\} \leq \inf_{\mathcal{C} \in \mathfrak{C}_{r_1}(U)} \sum \{(diam(C))^s \mid C \in \mathcal{C}\},$$

which means that $HB_{r_2}^s(U) \leq HB_{r_1}^s(U)$. Because of this, $\lim_{r \to 0} HB_r^s(U)$ exists for every set $U$, and this justifies the next definition.

**Definition 12.50.** *The* Hausdorff-Besicovitch outer measure $HB^s$ *is given by*

$$HB^s(U) = \lim_{r \to 0} HB_r^s(U)$$

*for every* $U \in \mathcal{P}(S)$.

**Theorem 12.51.** *Let* $(S, d)$ *be a metric space and let* $U$ *be a Borel set in this space. If* $s$ *and* $t$ *are two positive numbers such that* $s < t$ *and* $HB^s(U)$ *is finite, then* $HB^t(U) = 0$. *Further, if* $HB^t(U) > 0$, *then* $HB^s(U) = \infty$.

*Proof.* If $s < t$ and $\mathcal{C}$ is an $r$-cover of $U$, then

$$\begin{aligned}\sum \{(diam(C))^t \mid C \in \mathcal{C}\} &= \sum \{(diam(C))^{t-s}(diam(C))^s \mid C \in \mathcal{C}\} \\ &\leq r^{t-s} \sum \{(diam(C))^s \mid C \in \mathcal{C}\},\end{aligned}$$

which implies

$$HB_r^t(U) \leq r^{t-s} HB_r^s(U).$$

This, in turn, yields

$$HB^t(U) = \lim_{r \to 0} HB_r^t(U) \leq \lim_{r \to 0} r^{t-s} HB^s(U).$$

If $HB^s(U)$ is finite, then $HB^t(U) = 0$. On the other hand, if $HB^t(U) > 0$, the last inequality implies $HB^s(U) = \infty$. □

**Corollary 12.52.** *Let $(S, d)$ be a metric space and let $U$ be a Borel set. There exists a unique $s_0$ such that $0 \leq s_0 \leq \infty$ and*

$$HB^s(U) = \begin{cases} \infty & \text{if } s < s_0, \\ 0 & \text{if } s > s_0. \end{cases}$$

*Proof.* This statement follows immediately from Theorem 12.51 by defining $s_0$ as

$$s_0 = \inf\{s \in \mathbb{R}_{\geq 0} \mid HB^s(U) = 0\} = \sup\{s \in \mathbb{R}_{\geq 0} \mid HB^s(U) = \infty\}.$$

□

Corollary 12.52 justifies the following definition.

**Definition 12.53.** *Let $(S, d)$ be a metric space and let $U$ be a Borel set. The* Hausdorff-Besicovitch dimension of $U$ *is the number*

$$HBdim(U) = \sup\{s \in \mathbb{R}_{\geq 0} \mid HB^s(U) = \infty\}.$$

**Theorem 12.54.** *The Hausdorff-Besicovitch dimension is monotonic; that is, $U \subseteq U'$ implies $HBdim(U) \leq HBdim(U')$.*

*Proof.* If $U \subseteq U'$, then it is clear that $\mathfrak{C}_r(U') \subseteq \mathfrak{C}_r(U)$. Therefore, we have $HB^s_r(U) \leq HB^s_r(U')$, which implies $HB^s(U) \leq HB^s(U')$ for every $s \in \mathbb{R}_{\geq 0}$. This inequality yields $HBdim(U) \leq HBdim(U')$. □

**Theorem 12.55.** *If $\{U_n \mid n \in \mathbb{N}\}$ is a countable family of sets, then*

$$HBdim\left(\bigcup_{n\in\mathbb{N}} U_n\right) = \sup\{HBdim(U_n) \mid n \in \mathbb{N}\}.$$

*Proof.* By Theorem 12.55, we have $HBdim(U_n) \leq HBdim\left(\bigcup_{n\in\mathbb{N}} U_n\right)$, so $\sup\{HBdim(U_n) \mid n \in \mathbb{N}\} \leq HBdim\left(\bigcup_{n\in\mathbb{N}} U_n\right)$.

If $HBdim(U_n) < t$ for $n \in \mathbb{N}$, then $HB^t(U_n) = 0$, so $HB^t\left(\bigcup_{n\in\mathbb{N}} U_n\right) = 0$ since the Hausdorff-Besicovitch outer measure $HB^t$ is subadditive. Therefore, $HBdim\left(\bigcup_{n\in\mathbb{N}} U_n\right) < t$. This implies $HBdim\left(\bigcup_{n\in\mathbb{N}} U_n\right) \leq \sup\{HBdim(U_n) \mid n \in \mathbb{N}\}$, which gives the desired equality. □

*Example 12.56.* If $U = \{u\}$ is a singleton, then $HB^0(\{u\}) = 0$. Thus, $HBdim(\{x\}) = 0$. By Theorem 12.55, we have $HBdim(T) = 0$ for every countable set $T$.

*Example 12.57.* Let $f : [0, 1]^2 \longrightarrow \mathbb{R}$ be a function that is continuous and has bounded partial derivatives in the square $[0, 1]^2$ and let $S$ be the surface in $\mathbb{R}^2$ defined by $z = f(x, y)$. Under these conditions, there is a constant $k$ such that $|f(x', y') - f(x, y)| \leq k(|x' - x| + |y' - y|)$. We prove that $HBdim(S) = 2$.

Suppose that $S$ is covered by spheres of diameter $d_i$, $S \subseteq \bigcup\{B(x_i, \frac{d_i}{2}) \mid i \in I\}$. Then, the square $[0,1]^2$ is covered by disks of diameter $d_i$ and therefore $\sum_{i\in I} \frac{\pi d_i^2}{4} \geq 1$, which is equivalent to $\sum_{i\in I} d_i^2 \geq \frac{4}{\pi}$. Therefore, $HB^2(S) > 0$, so $HBdim(S) \geq 2$. Observe that, in this part of the argument, the regularity of $f$ played no role.

To prove the converse inequality, $HBdim(S) \leq 2$, we show that $HB^{2+\epsilon}(S) = 0$ for every $\epsilon > 0$; that is, $\lim_{r\to 0} HB_r^{2+\epsilon}(U) = 0$ for every $\epsilon > 0$.

Divide the square $[0,1]^2$ into $n^2$ squares of size $\frac{1}{n}$. Clearly, for any two pairs $(x, y)$ and $(x', y')$ located in the same small square, we have $|f(x', y') - f(x, y)| \leq \frac{2k}{n}$, which means that the portion of $S$ located above a small square can be enclosed in a cube of side $\frac{2k}{n}$ and therefore in a sphere of diameter $\frac{2\sqrt{3}k}{n}$. For the covering $\mathcal{C}$ that consists of these $n^2$ spheres, we have

$$\sum\{(diam(C))^{2+\epsilon} \mid C \in \mathcal{C}\} = n^2 \left(\frac{2\sqrt{3}k}{n}\right)^{2+\epsilon} = \frac{2\sqrt{3}k}{n^\epsilon}.$$

If $n$ is chosen such that $\frac{2\sqrt{3}k}{n} < r$, we have $HB_r^{2+\epsilon}(S) \leq \frac{2\sqrt{3}k}{n^\epsilon}$. Thus, $\lim_{r\to 0} HB_r^{2+\epsilon}(S) = 0$, so $HBdim(S) \leq 2$.

*Example 12.58.* We saw that the Cantor set $C$ is included in each of the sets $E_n$ that consists of $2^n$ closed intervals of length $\frac{1}{3^n}$. Thus, we have

$$HB^s_{\frac{1}{3^n}}(C) \leq \frac{2^n}{3^{sn}} = \left(\frac{2}{3^s}\right)^n$$

for every $n \geq 1$. If $\frac{2}{3^s} < 1$ (that is, if $s > \frac{\ln 2}{\ln 3}$), we have $\lim_{n\to\infty} HB^s_{\frac{1}{3^n}} = 0$. If $s < \frac{\ln 2}{\ln 3}$, then $\lim_{n\to\infty} HB^s_{\frac{1}{3^n}} = \infty$, so $HBdim(C) = \frac{\ln 2}{\ln 3}$.

**Theorem 12.59.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $T$ be a precompact set. We have $HBdim(T) \leq lbd(U)$.*

*Proof.* Suppose that $T$ can be covered by $n_T(r)$ sets of diameter $r$. By the definition of the outer measure $HB^s_r(U)$, we have

$$HB^s_r(U) \leq r^s n_T(r).$$

Since $HB^s(U) = \lim_{r\to 0} HB^s_r(U)$, if $HB^s(U) > 1$, then if $r$ is sufficiently small we have $HB^s_r(U) > 1$, so $\log HB^s_r(U) > 0$, which implies $s \log r + \log n_T(r) > 0$. Thus, if $r$ is sufficiently small, $s < \frac{n_T(r)}{\log \frac{1}{r}}$, so $s \leq lbd(U)$. This entails $HBdim(U) \leq lbd(U)$. □

The following statement is known as the *mass distribution principle* (see [48]).

**Theorem 12.60.** *Let $(S, d)$ be a metric space and let $\mu$ be a Carathéodory outer measure on $S$ such that there exist $s, r > 0$ such that $\mu(U) \leq c \cdot diam(U)^s$ for all $U \in \mathcal{P}(S)$ with $diam(U) \leq r$. Then, $HB^s(W) \leq \frac{\mu(W)}{c}$ and $s \leq HBdim(W) \leq lbd(W)$ for every precompact set $W \in \mathcal{P}(S)$ with $\mu(W) > 0$.*

*Proof.* Let $\{U_i \mid i \in I\}$ be a cover of $W$. We have

$$0 < \mu(W) \leq \mu\left(\bigcup_i U_i\right) \leq \sum_{i \in I} \mu(U_i) \leq c \sum_i (diam(U_i))^s,$$

so $\sum_i (diam(U_i))^s \geq \frac{\mu(W)}{c}$. Therefore, $HB_r^s(W) \geq \frac{\mu(W)}{c}$, which implies $HB^s(W) \geq \frac{\mu(W)}{c} > 0$. Consequently, $HBdim(W) \geq \frac{\mu(W)}{c} > 0$. □

## 12.8 Similarity Dimension

The notions of similarity and contraction between metric spaces were introduced in Definition 11.64.

**Definition 12.61.** *Let $\boldsymbol{r} = (r_1, \ldots, r_n)$ be a sequence of numbers such that $r_i \in (0,1)$ for $1 \leq i \leq n$ and let $(S,d)$ be a metric space.*

*An* iterative function system on $(S,d)$ *that realizes a sequence of ratios $\boldsymbol{r} = (r_1, \ldots, r_n)$ is a sequence of functions $\boldsymbol{f} = (f_1, \ldots, f_n)$, where $f_i : S \longrightarrow S$ is a contraction of ratio $r_i$ for $1 \leq i \leq n$.*

*A subset $T$ of $S$ is an* invariant set for the iterative function system $(f_1, \ldots, f_n)$ *if $T = \bigcup_{i=1}^n f_i(T)$.*

*Example 12.62.* Let $f_0, f_1 : [0,1] \longrightarrow [0,1]$ defined by

$$f_0(x) = \frac{x}{3} \text{ and } f_1(x) = \frac{x+2}{3}$$

for $x \in [0,1]$, which are contractions of ratio $\frac{1}{3}$.

The Cantor set $C$ is an invariant set for the iterative function system $\mathbf{f} = (f_0, f_1)$, as we have shown in Section 12.5.

**Lemma 12.63.** *Let $r_1, \ldots, r_n$ be $n$ numbers such that $r_i \in (0,1)$ for $1 \leq i \leq n$ and $n > 1$. There is a unique number $d$ such that*

$$r_1^d + r_2^d + \cdots + r_n^d = 1.$$

*Proof.* Define the function $\phi : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ by

$$\phi(x) = r_1^x + r_2^x + \cdots + r_n^x$$

for $x > 0$. Note that $\phi(0) = n$, $\lim_{x \to \infty} \phi(x) = 0$, and $\phi'(x) = r_1^x \ln r_1 + r_2^x \ln r_2 + r_n^x \ln r_n < 0$. Since $\phi'(x) < 0$, $\phi$ is a strictly decreasing function, so there exists a unique $d$ such that $\phi(d) = 1$. □

**Definition 12.64.** *Let $\boldsymbol{r} = (r_1, \ldots, r_n)$ be a sequence of ratios such that $r_i \in (0,1)$ for $1 \leq i \leq n$ and $n > 1$. The* dimension of $\boldsymbol{r}$ *is the number $d$, whose existence was proven in Lemma 12.63.*

Observe that if the sequence $\mathbf{r}$ has length 1, $\mathbf{r} = (r_1)$, then $r_1^d = 1$ implies $d = 0$.

*Example 12.65.* The dimension of the sequence $\mathbf{r} = (\frac{1}{3}, \frac{1}{3})$ is the solution of the equation

$$2 \cdot \left(\frac{1}{3}\right)^d = 1;$$

that is, $d = \frac{\log 2}{\log 3}$.

**Lemma 12.66.** *Let $(S, \mathcal{O}_d)$ be a complete topological metric space and let $\boldsymbol{f} = (f_1, \ldots, f_n)$ be an iterative function system that realizes a sequence of ratios $\boldsymbol{r} = (r_1, \ldots, r_n)$. The mapping $F : \mathcal{K}(S, \mathcal{O}_d) \longrightarrow \mathcal{K}(S, \mathcal{O}_d)$ defined on the Hausdorff metric hyperspace $(\mathcal{K}(S, \mathcal{O}_d), \delta)$ by*

$$F(U) = \bigcup_{i=1}^{n} f_i(U)$$

*is a contraction.*

*Proof.* We begin by observing that $F$ is well-defined. Indeed, since each contraction $f_i$ is continuous and the image of a compact set by a continuous function is compact (by Theorem 6.68), it follows that if $U$ is compact, then $F(U)$ is compact as the union of a finite collection of compact sets.

Next, we prove that $\delta(F(U), F(V)) \leq r\delta(U, V)$ for $r = \max_{0 \leq i \leq n-1} r_i < 1$.

Let $x \in F(U)$. There is $u \in U$ such that $x = f_i(u)$ for some $i$, $1 \leq i \leq n$. By the definition of $\delta$, there exists $v \in V$ such that $d(u, v) \leq \delta(U, V)$. Since $f_i$ is a contraction, we have $d(u, v) = d(f_i(v), f_i(u)) \leq r_i d(u, v) \leq rd(u, v) \leq r\delta(U, V)$, so $F(U) \subseteq C(F(V), r\delta(U, V))$. Similarly, $F(V) \subseteq C(F(U), r\delta(U, V))$, so $\delta(F(U), F(V)) \leq r\delta(U, V)$, which proves that $F$ is a contraction of the Hausdorff metric hyperspace $(\mathcal{K}(S, \mathcal{O}_d), \delta)$. □

**Theorem 12.67.** *Let $(S, \mathcal{O}_d)$ be a complete topological metric space and let $\boldsymbol{f} = (f_1, \ldots, f_n)$ be an iterative function system that realizes a sequence of ratios $\boldsymbol{r} = (r_1, \ldots, r_n)$. There exists a unique compact set $U$ that is an invariant set for $\boldsymbol{f}$.*

*Proof.* By Lemma 12.66, the mapping $F : \mathcal{K}(S, \mathcal{O}_d) \longrightarrow \mathcal{K}(S, \mathcal{O}_d)$ is a contraction. Therefore, by the Banach fixed point theorem (Theorem 11.70), $F$ has a fixed point in $\mathcal{K}(S, \mathcal{O}_d)$, which is an invariant set for $\mathbf{f}$. □

The unique compact set that is an invariant for an iterative function system $\mathbf{f}$ is usually referred to as the *attractor* of the system.

**Definition 12.68.** *Let $(S, \mathcal{O}_d)$ be a topological metric space and let $U$ be an invariant set of an iterative function system $\boldsymbol{f} = (f_1, \ldots, f_n)$ that realizes a sequence of ratios $\boldsymbol{r} = (r_1, \ldots, r_n)$.*

*The* similarity dimension of the pair $(U, \mathbf{f})$ *is the number SIMdim($\boldsymbol{f}$), which equals the dimension of $\boldsymbol{r}$.*

In principle, a set may be an invariant set for many iterative function systems.

*Example 12.69.* Let $p, q \in (0,1)$ such that $p + q \geq 1$ and let $f_0, f_1 : [0,1] \longrightarrow [0,1]$ be defined by $f_0(x) = px$ and $f_1(x) = qx + 1 - q$. Both $f_0$ and $f_1$ are contractions. The sequence $\mathbf{f} = (f_0, f_1)$ realizes the sequence of ratios $\mathbf{r} = (p, q)$ and we have $(0,1) = f_0(0,1) \cup f_1(0,1)$. The dimension of the pair $(U, \mathbf{r})$ is the number $d$ such that $p^d + q^d = 1$; this number depends on the values of $p$ and $q$.

**Theorem 12.70.** *Let $(S, \mathcal{O}_d)$ be a complete topological metric space, $\boldsymbol{f} = (f_1, \ldots, f_n)$ be an iterative function system that realizes a sequence of ratios $\boldsymbol{r} = (r_1, \ldots, r_n)$, and $U$ be the attractor of $\boldsymbol{f}$. Then, we have $HBdim(U) \leq SIMdim(\boldsymbol{f})$.*

*Proof.* Suppose that $r_1^d + r_2^d + \cdots + r_n^d = 1$, that is, $d$ is the dimension of $\mathbf{f}$. For a subset $T$ of $S$ denote the set $f_{i_1}(f_{i_2}(\cdots f_{i_p}(T)\cdots))$ by $f_{i_1 i_2 \cdots i_p}(T)$.

If $U$ is the attractor of $\mathbf{f}$, then

$$U = \bigcup \{f_{i_1 i_2 \cdots i_p}(U) \mid (i_1, i_2, \ldots, i_p) \in \mathbf{Seq}_p(\{1, \ldots, n\})\}.$$

This shows that the sets of the form $f_{i_1 i_2 \cdots i_p}(U)$ constitute a cover of $U$.

Since $f_{i_1}, f_{i_2}, \ldots, f_{i_p}$ are similarities of ratios $r_{i_1}, r_{i_2}, \ldots, r_{i_p}$, respectively, it follows that $diam(f_{i_1 i_2 \cdots i_p}(U)) \leq (r_{i_1} r_{i_2} \cdots r_{i_p}) diam(U)$. Thus,

$$\begin{aligned}
&\sum \{(diam(f_{i_1 i_2 \cdots i_p}(U)))^d \mid (i_1, i_2, \ldots, i_p) \in \mathbf{Seq}_p(\{1, \ldots, n\})\} \\
&\leq \sum \{r_{i_1}^d r_{i_2}^d \cdots r_{i_p}^d \, diam(U)^d \mid (i_1, i_2, \ldots, i_p) \in \mathbf{Seq}_p(\{1, \ldots, n\})\} \\
&= \left(\sum_{i_1} r_{i_1}^d\right)\left(\sum_{i_2} r_{i_2}^d\right) \cdots \left(\sum_{i_p} r_{i_p}^d\right) diam(U)^d \\
&= diam(U)^d.
\end{aligned}$$

For $r \in \mathbb{R}_{>0}$, choose $p$ such that $diam(f_{i_1 i_2 \cdots i_p}(U)) \leq (\max r_i)^p diam(U) < r$. This implies $HB_r^d(U) \leq diam(U)^d$, so $HB^d(U) = \lim_{r \to 0} HB_r^d(U) \leq diam(U)^d$. Thus, we have $HBdim(U) \leq d = SIMdim(\mathbf{f})$. □

The next statement involves an iterative function system $\mathbf{f} = (f_1, \ldots, f_m)$ acting on a closed subset $H$ of the metric space $(\mathbb{R}^n, d_2)$ such that each contraction $f_i$ satisfies the double inequality

$$b_i d_2(\mathbf{x}, \mathbf{y}) \leq d_2(f_i(\mathbf{x}), f_i(\mathbf{y})) \leq r_i d_2(\mathbf{x}, \mathbf{y})$$

for $1 \leq i \leq m$ and $\mathbf{x}, \mathbf{y} \in H$. Note that each of the functions $f_i$ is injective on the set $H$.

**Theorem 12.71.** *Let $\boldsymbol{f} = (f_1, \ldots, f_m)$ be an iterative function system on a closed subset $H$ of $\mathbb{R}^n$ that realizes a sequence of ratios $\boldsymbol{r} = (r_1, \ldots, r_m)$. Suppose that, for every $i$, $1 \leq i \leq m$, there exists $b_i \in (0,1)$ such that $d_2(f_i(\boldsymbol{x}), f_i(\boldsymbol{y})) \geq b_i d_2(\boldsymbol{x}, \boldsymbol{y})$ for $\boldsymbol{x}, \boldsymbol{y} \in H$.*

*If $U$ is the nonempty and compact attractor of $\boldsymbol{f}$ and $\{f_1(U), \ldots, f_m(U)\}$ is a partition of $U$, then $U$ is a totally disconnected set and $HBdim(U) \geq c$, where $c$ is the unique number such that $\sum_{i=1}^{n} b_i^c = 1$.*

*Proof.* Let

$$t = \min\{d_2(f_i(U), f_j(U) \mid 1 \leq i, j \leq m \text{ and } i \neq j\}.$$

Using the same notation as in the proof of Theorem 12.70, observe that the collection of sets

$$\{f_{i_1 \cdots i_p}(U) \mid (i_1, \ldots, i_p) \in \mathbf{Seq}(\{1, \ldots, m\})\}$$

is a sequential cover of the attractor $U$. Note also that all sets $f_{i_1 \cdots i_p}(U)$ are compact and therefore closed. Also, since each collection $\{f_{i_1 \cdots i_p}(U) \mid (i_1, \ldots, i_p) \in \mathbf{Seq}_p(\{1, \ldots, m\})$ is a partition of $U$, it follows that each of these sets is clopen in $U$. Thus, $U$ is totally disconnected.

Define

$$m(f_{i_1 i_2 \cdots i_p}(U)) = (b_{i_1} b_{i_2} \cdots b_{i_p})^c.$$

Note that

$$\begin{aligned}
\sum_{i=1}^{m} m(f_{i_1 i_2 \cdots i_p i}(U)) &= \sum_{i=1}^{m} (b_{i_1} b_{i_2} \cdots b_{i_p} b_i)^c \\
&= (b_{i_1} b_{i_2} \cdots b_{i_p})^c \sum_{i=1}^{m} b_i^c \\
&= (b_{i_1} b_{i_2} \cdots b_{i_p})^c = m(f_{i_1 i_2 \cdots i_p}(U)) \\
&= m\left( \bigcup_{i=1}^{m} f_{i_1 i_2 \cdots i_p i}(U) \right).
\end{aligned}$$

For $x \in U$, there is a unique sequence $(i_1, i_2, \ldots) \in \mathbf{Seq}_\infty(\{1, \ldots, m\})$ such that $x \in U_{i_1 \cdots i_k}$ for every $k \geq 1$. Observe also that

$$U \supseteq U_{i_1} \supseteq U_{i_1 i_2} \supseteq \cdots \supseteq U_{i_1 i_2 \cdots i_k} \supseteq \cdots .$$

Consider the decreasing sequence $1 > b_{i_1} > b_{i_1} b_{i_2} > \cdots > b_{i_1} \cdots b_{i_n} > \cdots$. If $0 < r < t$, let $k$ be the least number $j$ such that $\frac{r}{t} \geq b_{i_1} \cdots b_{i_j}$. We have

$$b_{i_1} \cdots b_{i_{k-1}} > \frac{r}{t} \geq b_{i_1} \cdots b_{i_k}$$

so $m(U_{i_1 \cdots i_{k-1}}) > \frac{r^c}{t^c} \geq m(U_{i_1 \cdots i_k})$.

Let $(i'_1, \ldots, i'_k)$ be a sequence distinct from $(i_1, \ldots, i_k)$. If $\ell$ is the least integer such that $i'_\ell \neq i_\ell$, then $U_{i'_1 \ldots i'_\ell} \subseteq U_{i'_\ell}$ and $U_{i_1 \ldots i_\ell} \subseteq U_{i_\ell}$. Since $U_{i_\ell}$ and $U_{i_{\ell'}}$ are disjoint and separated by $t$, it follows that the sets $U_{i_1 \ldots i_k}$ and $U_{i'_1 \ldots i'_k}$ are disjoint and separated by at least $b_{i_1} \cdots b_{i_\ell} t > r$. Thus, $U \cap B(x,r) \subset U_{i_1 \cdots i_k}$, so

$$m(U \cap B(x,r)) \leq m(U_{i_1 \cdots i_k}) = (b_{i_1} \cdots b_{i_k})^c \leq \left(\frac{r}{t}\right)^c.$$

If $U \cap W \neq \emptyset$, then $W \subset B(x,r)$ for some $x \in U$ with $r = \mathit{diam}(W)$. Thus, $m(W) \leq \frac{\mathit{diam}(W)}{t^c}$, so $HB^c(U) > 0$ and $HBdim(U) \geq c$. $\square$

## Exercises and Supplements

1. Let $Q_n(\ell)$ be an $n$-dimensional cube in $\mathbb{R}^n$. Prove that:
   a) there are $\binom{n}{k} \cdot 2^{n-k}$ $k$-dimensional faces of $Q_n(\ell)$;
   b) the total number of faces of $Q_n(\ell)$ is $\sum_{k=1}^{n} \binom{n}{k} \cdot 2^{n-k} = 3^n - 1$.
2. Let $T$ be a subset of $\mathbb{R}$. Prove that $T$ is zero-dimensional if and only if it contains no interval.
3. Prove that the Cantor set $C$ is totally disconnected.
   **Hint:** Suppose that $a < b$ and $b$ belongs to the connected components $K_a$ of $a$. By Example 6.82, this implies $[a,b] \subseteq K_a \subseteq C$, which leads to a contradiction.
4. Prove the following extension of Example 12.42. If $T = \{0\} \cup \{\frac{1}{n^a} \mid n \geq 1\}$, then $(\mathbf{T}) = \frac{\mathbf{1}}{\mathbf{1+a}}$.
5. Let $(S, \mathcal{O}_d)$ be a compact topological metric space, $x \in S$, and let $H$ be a closed set in $(S, \mathcal{O})$. Prove that if the sets $\{x\}$ and $\{y\}$ are separated by a closed set $K_{xy}$ with $\mathit{ind}(K_{xy}) \leq n-1$ for every $y \in H$, then $\{x\}$ is separated from $H$ by a closed set $K$ with $\mathit{ind}(K) \leq n-1$.
6. Prove that every zero-dimensional separable topological space $(S, \mathcal{O})$ is homeomorphic to a subspace of the Cantor set.
   **Hint:** By the separability of $(S, \mathcal{O})$ and by Theorem 12.6, $(S, \mathcal{O})$ has a countable basis $\{B_0, B_1, \ldots, B_n, \ldots\}$ that consists of clopen sets. Consider the mapping $f : S \longrightarrow \mathbf{Seq}_\infty(\{0,1\})$ defined by $f(x) = (b_0, b_1, \ldots)$, where $b_i = I_{B_i}(x)$ for $i \in \mathbb{N}$.
7. Let $T$ be a subset of $\mathbb{R}^n$. A function $f : T \longrightarrow \mathbb{R}^n$ satisfies the *Hölder condition of exponent* $\alpha$ if there is a constant $k$ such that $|f(\mathbf{x}) - f(\mathbf{y})| \leq k|\mathbf{x} - \mathbf{y}|^\alpha$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Prove that

$$HB^{\frac{s}{\alpha}}(f(T)) \leq k^{\frac{s}{\alpha}} HB^s(T).$$

   **Solution:** If $C \subseteq \mathbb{R}^n$ is a set of diameter $\mathit{diam}(C)$, then $f(C)$, the image of $C$ under $f$, has a diameter no larger than $k(\mathit{diam}(C))^\alpha$. Therefore, if $\mathcal{C}$ is an $r$-cover of $T$, then $\{f(T \cap C) \mid C \in \mathcal{C}\}$ is a $kr^\alpha$-cover of $f(T)$. Therefore,

$$\sum\{(diam(T \cap C))^{\frac{s}{\alpha}} \mid C \in \mathcal{C}\} \leq \sum\{(k(diam(C))^{\alpha})^{\frac{s}{\alpha}} \mid C \in C\}$$
$$= k^{\frac{s}{\alpha}} \sum\{(diam(C))^{s} \mid C \in \mathcal{C}\},$$

which implies $HB_r^{\frac{s}{\alpha}}(T) \leq k^{\frac{s}{\alpha}} HB_r^s(T)$. Since $\lim_{r \to 0} kr^{\alpha} = 0$, we have $HB^{\frac{s}{\alpha}}(f(T)) \leq k^{\frac{s}{\alpha}} HB^s(T)$.

8. Consider the ultrametric space $(\mathbf{Seq}_{\infty}(\{0,1\}), d_{\phi})$ introduced in Supplement 46 of Chapter 10, where $\phi(\mathbf{u}) = 2^{-|\mathbf{u}|}$ for $\mathbf{u} \in \mathbf{Seq}(\{0,1\})$. Prove that if $(S, \mathcal{O}_d)$ is a separable topological metric space such that $ind(S) = 0$, then $S$ is homeomorphic to a subspace of the topological metric space $(\mathbf{Seq}_{\infty}\{0,1\}, \mathcal{O}_{d_{\phi}})$.

   **Solution:** Since $ind(S) = 0$, there exists a basis $\mathcal{B}_0$ for $\mathcal{O}_d$ that consists of clopen sets (by Theorem 12.6). Further, since $(S, \mathcal{O}_d)$ is countable, there exists a basis $\mathcal{B} \subseteq \mathcal{B}_0$ that is countable. Let $\mathcal{B} = \{B_0, B_1, \ldots\}$.

   If $\mathbf{s} = (s_0, s_1, \ldots, s_{p-1}) \in \mathbf{Seq}_p(\{0,1\})$, let $B(\mathbf{s})$ be the clopen set $B(\mathbf{s}) = B_0^{s_0} \cap B_1^{s_1} \cap \cdots \cap B_{p-1}^{s_1}$.

   Define the mapping $h : S \longrightarrow \mathbf{Seq}_{\infty}(\{0,1\})$ by $h(x) = (s_0, s_1, \ldots)$, where $s_i = 1$ if $x \in B_i$ and $s_i = 0$ otherwise for $x \in S$. Thus, $\mathbf{s}$ is a prefix of $h(x)$ if and only if $x \in B(\mathbf{s})$. The mapping $h$ is injective. Indeed, suppose that $x \neq y$. Since $S - \{y\}$ is an open set containing $x$, there exists $i$ with $x \in B_i \subseteq S - \{y\}$, which implies $(h(x))_i = 1$ and $(h(y))_i = 0$, so $h(x) \neq h(y)$. Thus, $h$ is a bijection between $S$ and $h(S)$.

   In Exercise 5 of Chapter 11, we saw that the collection $\{P_{\mathbf{u}} \mid \mathbf{u} \in \mathbf{Seq}(\{0,1\})\}$ is a basis for $\mathbf{Seq}_{\infty}(\{0,1\})$ and $h^{-1}(P_{\mathbf{u}}) = B(\mathbf{u})$, so $h^{-1} : h(S) \longrightarrow S$ is continuous.

   Note that $h(U_i) = h(S) \cap \{0,1\}^{i-1}\mathbf{Seq}_{\infty}(\{0,1\})$ is open in $\mathbf{Seq}_{\infty}(\{0,1\})$ for every $i \in \mathbb{N}$, so $h^{-1}$ is continuous. Thus, $h$ is a homeomorphism of $S$ into $h(S)$.

9. Let $\mathbf{f} = (f_1, \ldots, f_m)$ be an iterative function system on $\mathbb{R}^n$ that realizes the sequence $(r, \ldots, r)$ with $r \in (0,1)$ and let $H$ be a nonempty compact set in $\mathbb{R}^n$. Prove that if $U$ is the attractor of $\mathbf{f}$, then

$$\delta(H, U) \leq \frac{1}{1-r} \delta\left(H, F(H)\right),$$

   where $\delta$ is the metric of the Hausdorff hyperspace of compact subsets and $F(T) = \bigcup_{i=1}^{n} f_i(T)$ for $T \in \mathcal{P}(S)$.

   **Solution:** In the proof of Lemma 12.66, we saw that $\delta(F(H), F(U)) \leq r\delta(H, U)$. This allows us to write

$$\begin{aligned} \delta(H, U) &\leq \delta(H, F(H)) + \delta(F(H), U) \\ &= \delta(H, F(H)) + \delta(F(H), F(U)) \\ &\leq \delta(H, F(H)) + r\delta(H, U), \end{aligned}$$

   which implies the desired inequality.

10. Consider the contractions $f_0, f_1 : \mathbb{R} \longrightarrow \mathbb{R}$ defined by $f_0(x) = rx$ and $f_1(x) = rx + 1 - r$ for $x \in \mathbb{R}$, where $r \in (0, 1)$. Find the attractor of the iterative function system $\mathbf{f} = (f_0, f_1)$.
11. Let $(S, \mathcal{O}_d)$ be a compact topological metric space. Prove that $cov(S) \leq n$ if and only if for every $\epsilon > 0$ there is an open cover $\mathcal{C}$ of $S$ with $ord(\mathcal{C}) \leq n$ and $\sup\{diam(C) \mid C \in \mathcal{C}\} < \epsilon$.

    **Solution:** The set of open spheres $\{C(x, \epsilon/2) \mid x \in S\}$ is an open cover that has a refinement with order not greater than $n$; the diameter of each of the sets of the refinement is less than $\epsilon$.

    Conversely, suppose that for every $\epsilon > 0$ there is an open cover $\mathcal{C}$ of $S$ with $ord(\mathcal{C}) \leq n$ and $\sup\{diam(C) \mid C \in \mathcal{C}\} < \epsilon$.

    Let $\mathcal{D}$ be a finite open cover of $S$. By Lebesgue's Lemma (Theorem 11.19) there exists $r > 0$ such that for every subset $U$ of $S$ with $diam(U) < r$ there is a set $L \in \mathcal{D}$ such that $U \subseteq L$.

    Let $\mathcal{C}'$ be an open cover of $S$ with order not greater than $n$ and such that $\sup\{diam(C') \mid C' \in \mathcal{C}'\} < \min\{\epsilon, r\}$. Then $\mathcal{C}'$ is a refinement of $\mathcal{D}$, so $cov(S) \leq n$.
12. Prove that if $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is an isometry, then $HB^{\frac{s}{\alpha}}(f(T)) = HB^s(T)$.
13. Let $F$ be a finite set of a metric space. Prove that $HB^0(F) = |F|$.
14. A useful variant of the Hausdorff-Besicovitch outer measure can be defined by restricting the $r$-covers to closed spheres of radius no greater than $r$. Let $\mathfrak{B}_r(U)$ be the set of all countable covers of a subset $U$ of a metric space $(S, d)$ that consist of closed spheres of radius no greater than $r$. Since $\mathfrak{B}_r(U) \subseteq \mathfrak{C}_r(U)$, it is clear that $HB^s_r(U) \leq HB^{'s}_r(U)$, where

$$HB^{'s}_r(U) = \inf_{\mathcal{C} \in \mathfrak{B}_r(U)} \sum \{(diam(C))^s \mid C \in \mathcal{C}\}.$$

    Prove that:
    a) $HB^{'s}_r(U) \leq 2^s HB^s_r(U)$,
    b) $HB^s U \leq HB^{'s} U \leq 2^s HB^s U$, where $HB^{'s}(U) = \lim_{r \to 0} HB^{'s}_r(U)$, and
    c) $HBdim(U) = HBdim'(U)$, where $HBdim'(U) = \sup\{s \in \mathbb{R}_{\geq 0} \mid HB^{'s}(U) = \infty\}$

    for every Borel subset $U$ of $S$.
15. Prove that if $U$ is a subset of $\mathbb{R}^n$ such that $\mathbf{I}(U) \neq \emptyset$, then $HBdim(U) = n$.

Let $(S, d)$ be a metric space, $s$ and $r$ be two numbers in $\mathbb{R}_{>0}$, $U$ be a subset of $S$, and

$$P^s_r(U) = \sup \left\{ \sum_i diam(B_i)^s \mid B_i \in \mathcal{B}_r(U) \right\},$$

where $\mathcal{B}_r(U)$ is the collection of disjoint closed spheres centered in $U$ and having diameter not larger than $r$. Observe that $\lim_{r \to 0} P^s_r(U)$ exists because $P^s_r(U)$ decreases when $r$ decreases. Let $P^s(U) = \lim_{r \to 0} P^s_r(U)$.

16. Let $(S, \mathcal{O}_d)$ be a topological metric space. Define $PK^s(U)$ as the outer measure obtained by Method I starting from the function $P^s$,

$$PK^s(U) = \inf_{\mathcal{C} \in \mathfrak{C}_r(U)} \sum \{P^s(C) \mid C \in \mathcal{C}\},$$

where $\mathfrak{C}_r(U)$ is the collection of all countable $r$-covers for a set $U$.

a) Prove that if $U$ is a Borel set in $(S, \mathcal{O}_d)$, $0 < s < t$, and $PK^s(U)$ is finite, then $PK^t(U) = 0$. Further, prove that if $PK^t(U) > 0$, then $PK^s(U) = \infty$.

b) The *packing dimension* of $U$ is defined as

$$PKdim(U) = \sup\{s \mid PK^s(U) = \infty\}.$$

Prove that $HBdim(U) \leq PKdim(U)$ for any Borel subset $U$ of $\mathbb{R}^n$.

## Bibliographical Comments

The first monograph dedicated to dimension theory is the book by Hurewicz and Wallman [71]. A topology source with a substantial presentation of topological dimensions is [47]. The literature dedicated to fractals has several excellent references for dimension theory [44, 45, 48]. Supplement 9 appears in the last reference. Example 12.57 is from [13], where an interesting connection between entropy and the Hausdorff-Besicovitch dimension is discussed.

# 13

# Clustering

## 13.1 Introduction

Clustering is the process of grouping together objects that are similar. The groups formed by clustering are referred to as *clusters*. Similarity between objects that belong to a set $S$ is usually measured using a dissimilarity $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ that is definite (see Section 10.2). This means that $d(x, y) = 0$ if and only if $x = y$ and $d(x, y) = d(y, x)$ for every $x, y \in S$. Two objects $x$ and $y$ are similar if the value of $d(x, y)$ is small; what "small" means depends on the context of the problem.

Clustering can be regarded as a special type of classification, where the clusters serve as classes of objects. It is a widely used data mining activity with multiple applications in a variety of scientific activities ranging from biology and astronomy to economics and sociology.

There are several points of view for examining clustering techniques. We follow here the taxonomy of clustering presented in [73].

Clustering may or may not be *exclusive*, where an exclusive clustering technique yields clusters that are disjoint, while a nonexclusive technique produces overlapping clusters. From an algebraic point of view, an exclusive clustering algorithm generates a partition of the set of objects, and most clustering algorithms fit in this category.

Clustering may be *intrinsic* or *extrinsic*. Intrinsic clustering is an unsupervised activity that is based only on the dissimilarities between the objects to be clustered. Most clustering algorithms fall into this category. Extrinsic clustering relies on information provided by an external source that prescribes, for example, which objects should be clustered together and which should not.

Finally, clustering may be *hierarchical* or *partitional*.

In hierarchical clustering algorithms, a sequence of partitions is constructed. In *hierarchical agglomerative algorithms* this sequence is increasing and it begins with the least partition of the set of objects whose blocks consist of single objects; as the clustering progresses, certain clusters are fused together. As a result, an agglomerative clustering is a chain of partitions on

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_13, © Springer-Verlag London Limited 2008

the set of objects that begins with the least partition $\alpha_S$ of the set of objects $S$ and ends with the largest partition $\omega_S$. In a *hierarchical divisive algorithm*, the sequence of partitions is decreasing. Its first member is the one-block partition $\omega_S$, and each partitions is built by subdividing the blocks of the previous partition.

Partitional clustering creates a partition of the set of objects whose blocks are the clusters such that objects in a cluster are more similar to each other than to objects that belong to different clusters. A typical representative algorithm is the $k$-means algorithm and its many extensions.

Our presentation is organized around the last dichotomy. We start with a class of hierarchical agglomerative algorithms. This is continued with a discussion of the $k$-means algorithm, a representative of partitional algorithms. Then, we continue with a discussion of certain limitations of clustering centered around Kleinberg's impossibility theorem. We conclude with an evaluation of clustering quality.

## 13.2 Hierarchical Clustering

Hierarchical clustering is a recursive process that begins with a metric space of objects $(S, d)$ and results in a chain of partitions of the set of objects. In each of the partitions, similar objects belong to the same block and objects that belong to distinct blocks tend to be dissimilar.

In agglomerative hierarchical clustering, the construction of this chain begins with the unit partition $\pi^1 = \alpha_S$. If the partition constructed at step $k$ is

$$\pi^k = \{U_1^k, \ldots, U_{m_k}^k\},$$

then two distinct blocks $U_p^k$ and $U_q^k$ of this partition are selected using a *selection criterion*. These blocks are fused and a new partition

$$\pi^{k+1} = \{U_1^k, \ldots, U_{p-1}^k, U_{p+1}^k, \ldots, U_{q-1}^k, U_{q+1}^k, \ldots, U_p^k \cup U_q^k\}$$

is formed. Clearly, we have $\pi^k \prec \pi^{k+1}$. The process must end because the poset $(PART(S), \leq)$ is of finite height. The algorithm halts when the one-block partition $\omega_S$ is reached.

As we saw in Theorem 10.28, the chain of partitions $\pi^1, \pi^2, \ldots$ generates a hierarchy on the set $S$. Therefore, all tools developed for hierarchies, including the notion of a dendrogram, can be used for hierarchical algorithms.

When data to be clustered are numerical (that is, when $S \subseteq \mathbb{R}^n$), we can define the *centroid* of a nonempty subset $U$ of $S$ as:

$$\mathbf{c}_U = \frac{1}{|U|} \sum \{\mathbf{o} | \mathbf{o} \in U\}.$$

If $\pi = \{U_1, \ldots, U_m\}$ is a partition of $S$, then the *sum of the squared errors* of $\pi$ is the number

$$sse(\pi) = \sum_{i=1}^{m} \sum \{d^2(\mathbf{o}, \mathbf{c}_{U_i}) | \mathbf{o} \in U_i\}, \tag{13.1}$$

where $d$ is the Euclidean distance in $\mathbb{R}^n$.

If two blocks $U$ and $V$ of a partition $\pi$ are fused into a new block $W$ to yield a new partition $\pi'$ that covers $\pi$, then the variation of the sum of squared errors is given by

$$\begin{aligned} sse(\pi') - sse(\pi) = & \sum \{d^2(\mathbf{o}, \mathbf{c}_W) | \mathbf{o} \in U \cup V\} \\ & - \sum \{d^2(\mathbf{o}, \mathbf{c}_U) | \mathbf{o} \in U\} - \sum \{d^2(\mathbf{o}, \mathbf{c}_V) | \mathbf{o} \in V\}. \end{aligned}$$

The centroid of the new cluster $W$ is given by

$$\begin{aligned} \mathbf{c}_W &= \frac{1}{|W|} \sum \{\mathbf{o} | \mathbf{o} \in W\} \\ &= \frac{|U|}{|W|} \mathbf{c}_U + \frac{|V|}{|W|} \mathbf{c}_V. \end{aligned}$$

This allows us to evaluate the increase in the sum of squared errors:

$$\begin{aligned} sse(\pi') - sse(\pi) = & \sum \{d^2(\mathbf{o}, \mathbf{c}_W) \mid \mathbf{o} \in U \cup V\} \\ & - \sum \{d^2(\mathbf{o}, \mathbf{c}_U) \mid \mathbf{o} \in U\} - \sum \{d^2(\mathbf{o}, \mathbf{c}_V) \mid \mathbf{o} \in V\} \\ = & \sum \{d^2(\mathbf{o}, \mathbf{c}_W) - d^2(\mathbf{o}, \mathbf{c}_U) \mid \mathbf{o} \in U\} \\ & + \sum \{d^2(\mathbf{o}, \mathbf{c}_W) - d^2(\mathbf{o}, \mathbf{c}_V) \mid \mathbf{o} \in V\}. \end{aligned}$$

Observe that:

$$\begin{aligned} & \sum \{d^2(\mathbf{o}, \mathbf{c}_W) - d^2(\mathbf{o}, \mathbf{c}_U) \mid \mathbf{o} \in U\} \\ & = \sum_{\mathbf{o} \in U} ((\mathbf{o} - \mathbf{c}_W)(\mathbf{o} - \mathbf{c}_W) - (\mathbf{o} - \mathbf{c}_U)(\mathbf{o} - \mathbf{c}_U)) \\ & = |U|(\mathbf{c}_W^2 - \mathbf{c}_U^2) + 2(\mathbf{c}_U - \mathbf{c}_W) \sum_{\mathbf{o} \in U} \mathbf{o} \\ & = |U|(\mathbf{c}_W^2 - \mathbf{c}_U^2) + 2|U|(\mathbf{c}_U - \mathbf{c}_W)\mathbf{c}_U \\ & = (\mathbf{c}_W - \mathbf{c}_U)\,(|U|(\mathbf{c}_W + \mathbf{c}_U) - 2|U|\mathbf{c}_U) \\ & = |U|(\mathbf{c}_W - \mathbf{c}_U)^2. \end{aligned}$$

Using the equality $\mathbf{c}_W - \mathbf{c}_U = \frac{|U|}{|W|}\mathbf{c}_U + \frac{|V|}{|W|}\mathbf{c}_V - \mathbf{c}_U = \frac{|V|}{|W|}(\mathbf{c}_V - \mathbf{c}_U)$, we obtain $\sum \{d^2(\mathbf{o}, \mathbf{c}_W) - d^2(\mathbf{o}, \mathbf{c}_U) \mid \mathbf{o} \in U\} = \frac{|U||V|^2}{|W|^2}(\mathbf{c}_V - \mathbf{c}_U)^2$.

Similarly, we have

$$\sum \{d^2(\mathbf{o}, \mathbf{c}_W) - d^2(\mathbf{o}, \mathbf{c}_V) \mid \mathbf{o} \in V\} = \frac{|U|^2|V|}{|W|^2}(\mathbf{c}_V - \mathbf{c}_U)^2,$$

so,

$$sse(\pi') - sse(\pi) = \frac{|U||V|}{|W|}\left(\mathbf{c}_V - \mathbf{c}_U\right)^2. \tag{13.2}$$

The dissimilarity between two clusters $U$ and $V$ can be defined using one of the following real-valued, two-argument functions defined on the set of subsets of $S$:

$$\begin{aligned} sl(U,V) &= \min\{d(u,v)|u \in U, v \in V\};\\ cl(U,V) &= \max\{d(u,v)|u \in U, v \in V\};\\ gav(U,V) &= \frac{\sum\{d(u,v)|u \in U, v \in V\}}{|U| \cdot |V|};\\ cen(U,V) &= (\mathbf{c}_U - \mathbf{c}_V)^2;\\ ward(U,V) &= \frac{|U||V|}{|U| + |V|}\left(\mathbf{c}_V - \mathbf{c}_U\right)^2. \end{aligned}$$

The names of the functions *sl*, *cl*, *gav*, and *cen* defined above are acronyms of the terms "single link", "complete link", "group average", and "centroid", respectively. They are linked to variants of the hierarchical clustering algorithms that we discuss in later. Note that in the case of the *ward* function the value equals the increase in the sum of the square errors when the clusters $U, V$ are replaced with their union.

### 13.2.1 Matrix-Based Hierarchical Clustering

The specific selection criterion for fusing blocks defines the clustering algorithm. All algorithms store the dissimilarities between the current clusters $\pi^k = \{U_1^k, \ldots, U_{m_k}^k\}$ in an $m_k \times m_k$-matrix $D^k = (d_{ij}^k)$, where $d_{ij}^k$ is the dissimilarity between the clusters $U_i^k$ and $U_j^k$. As new clusters are created by merging two existing clusters, the distance matrix must be adjusted to reflect the dissimilarities between the new cluster and existing clusters.

The general form of the algorithm is as follows.

**Algorithm 13.1 (Matrix Agglomerative Clustering)**
**Input:** the initial dissimilarity matrix $D^1$.
**Output:** the cluster hierarchy on the set of objects $S$, where $|S| = n$;
**Method:**
$k = 1$;
initialize clustering: $\pi^1 = \alpha_S$;
**while** ($\pi^k$ contains more than one block) **do**
merge a pair of two of the closest clusters;
output new cluster;
$k++$;
compute the dissimilarity matrix $D^k$;
**endwhile**

To evaluate the space and time complexity of hierarchical clustering, note that the algorithm must handle the matrix of the dissimilarities between objects, and this is a symmetric $n \times n$-matrix having all elements on its main diagonal equal to 0; in other words, the algorithm needs to store $\frac{n(n-1)}{2}$ numbers. To keep track of the clusters, an extra space that does not exceed $n-1$ is required. Thus, the total space required is $O(n^2)$.

The time complexity of agglomerative clustering algorithms has been evaluated in [84]; the proposed implementation requires a heap that contains the pairwise distances between clusters and therefore has a size of $n^2$.

Observe that the while loop is performed $n$ times as each execution reduces the number of clusters by 1. The initial construction of the heap requires a time of $O(n^2 \log n^2) = O(n^2 \log n)$. Then, each operation inside the loop requires no more than $O(\log n^2) = O(\log n)$ (because the heap has size $n^2$). Thus, we conclude that the time complexity is $O(n^2 \log n)$.

The computation of the dissimilarity between a new cluster and existing clusters is described next.

**Theorem 13.2.** *Let $U$ and $V$ be two clusters of the clustering $\pi$ that are joined into a new cluster $W$. Then, if $Q \in \pi - \{U, V\}$, we have*

$$\begin{aligned}
sl(W,Q) &= \frac{1}{2}sl(U,Q) + \frac{1}{2}sl(V,Q) - \frac{1}{2}\Big|sl(U,Q) - sl(V,Q)\Big|;\\
cl(W,Q) &= \frac{1}{2}cl(U,Q) + \frac{1}{2}cl(V,Q) + \frac{1}{2}\Big|cl(U,Q) - cl(V,Q)\Big|;\\
gav(W,Q) &= \frac{|U|}{|U|+|V|}gav(U,Q) + \frac{|V|}{|U|+|V|}gav(V,Q);\\
cen(W,Q) &= \frac{|U|}{|U|+|V|}cen(U,Q) + \frac{|V|}{|U|+|V|}cen(V,Q)\\
&\quad - \frac{|U||V|}{(|U|+|V|)^2}cen(U,V);\\
ward(W,Q) &= \frac{|U|+|Q|}{|U|+|V|+|Q|}ward(U,Q) + \frac{|V|+|Q|}{|U|+|V|+|Q|}ward(V,Q)\\
&\quad - \frac{|Q|}{|U|+|V|+|Q|}ward(U,V).
\end{aligned}$$

*Proof.* The first two equalities follow from the fact that

$$\begin{aligned}
\min\{a,b\} &= \frac{1}{2}(a+b) - \frac{1}{2}|a-b|,\\
\max\{a,b\} &= \frac{1}{2}(a+b) + \frac{1}{2}|a-b|,
\end{aligned}$$

for every $a, b \in \mathbb{R}$.

For the third equality, we have

$$\begin{aligned} gav(W,Q) &= \frac{\sum\{d(w,q)|w \in W, q \in Q\}}{|W| \cdot |Q|} \\ &= \frac{\sum\{d(u,q)|u \in U, q \in Q\}}{|W| \cdot |Q|} + \frac{\sum\{d(v,q)|v \in V, q \in Q\}}{|W| \cdot |Q|} \\ &= \frac{|U|}{|W|} \frac{\sum\{d(u,q)|u \in U, q \in Q\}}{|U| \cdot |Q|} + \frac{|V|}{|W|} \frac{\sum\{d(v,q)|v \in V, q \in Q\}}{|V| \cdot |Q|} \\ &= \frac{|U|}{|U|+|V|} gav(U,Q) + \frac{|V|}{|U|+|V|} gav(V,Q). \end{aligned}$$

The equality involving the function *cen* is immediate. The last equality can be easily translated into

$$\begin{aligned} &\frac{|Q||W|}{|Q|+|W|} \left(\mathbf{c}_Q - \mathbf{c}_W\right)^2 \\ &= \frac{|U|+|Q|}{|U|+|V|+|Q|} \frac{|U||Q|}{|U|+|Q|} \left(\mathbf{c}_Q - \mathbf{c}_U\right)^2 \\ &\quad + \frac{|V|+|Q|}{|U|+|V|+|Q|} \frac{|V||Q|}{|V|+|Q|} \left(\mathbf{c}_Q - \mathbf{c}_V\right)^2 \\ &\quad - \frac{|Q|}{|U|+|V|+|Q|} \frac{|U||V|}{|U|+|V|} \left(\mathbf{c}_V - \mathbf{c}_U\right)^2, \end{aligned}$$

which can be verified replacing $|W| = |U| + |V|$ and $\mathbf{c}_W = \frac{|U|}{|W|}\mathbf{c}_U + \frac{|V|}{|W|}\mathbf{c}_V$. □

The equalities contained by Theorem 13.2 are often presented as a single equality involving several coefficients.

**Corollary 13.3 (The Lance-Williams Formula).** *Let $U$ and $V$ be two clusters of the clustering $\pi$ that are joined into a new cluster $W$. Then, if $Q \in \pi - \{U, V\}$, the dissimilarity between $W$ and $Q$ can be expressed as*

$$d(W,Q) = a_U d(U,Q) + a_V d(V,Q) + bd(U,V) + c|d(U,Q) - d(V,Q)|,$$

*where the coefficients $a_U, a_V, b, c$ are given by the following table:*

| Function | $a_U$ | $a_V$ | $b$ | $c$ |
|---|---|---|---|---|
| *sl* | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $-\frac{1}{2}$ |
| *cl* | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{2}$ |
| *gav* | $\frac{\lvert U\rvert}{\lvert U\rvert+\lvert V\rvert}$ | $\frac{\lvert V\rvert}{\lvert U\rvert+\lvert V\rvert}$ | 0 | 0 |
| *cen* | $\frac{\lvert U\rvert}{\lvert U\rvert+\lvert V\rvert}$ | $\frac{\lvert V\rvert}{\lvert U\rvert+\lvert V\rvert}$ | $-\frac{\lvert U\rvert\lvert V\rvert}{(\lvert U\rvert+\lvert V\rvert)^2}$ | 0 |
| *ward* | $\frac{\lvert U\rvert+\lvert Q\rvert}{\lvert U\rvert+\lvert V\rvert+\lvert Q\rvert}$ | $\frac{\lvert V\rvert+\lvert Q\rvert}{\lvert U\rvert+\lvert V\rvert+\lvert Q\rvert}$ | $-\frac{\lvert Q\rvert}{\lvert U\rvert+\lvert V\rvert+\lvert Q\rvert}$ | 0 |

*Proof.* This statement is an immediate consequence of Theorem 13.2. □

The variant of the algorithm that makes use of the function *sl* is known as the *single-link* clustering. It tends to favor elongated clusters.

*Example 13.4.* We use single-link clustering for the metric space $(S, d_1)$, where $S \subseteq \mathbb{R}^2$ consists of seven objects, $S = \{\mathbf{o}_1, \ldots, \mathbf{o}_7\}$ (see Figure 13.1).

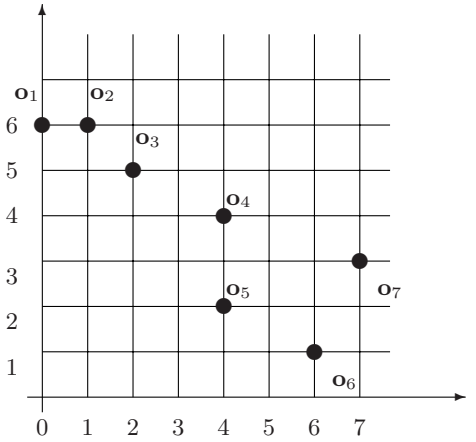



**Fig. 13.1.** Set of seven points in $\mathbb{R}^2$.

The distances $d_1(\mathbf{o}_i, \mathbf{o}_j)$ for $1 \leq i, j \leq 7$ between the objects of $S$ are specified by the $7 \times 7$ matrix

$$D^1 = \begin{pmatrix} 0 & 1 & 3 & 6 & 8 & 11 & 10 \\ 1 & 0 & 2 & 5 & 7 & 10 & 9 \\ 3 & 2 & 0 & 3 & 5 & 8 & 7 \\ 6 & 5 & 3 & 0 & 2 & 5 & 4 \\ 8 & 7 & 5 & 2 & 0 & 3 & 4 \\ 11 & 10 & 8 & 5 & 3 & 0 & 3 \\ 10 & 9 & 7 & 4 & 4 & 3 & 0 \end{pmatrix}.$$

We apply the hierarchical clustering algorithm using the single-link variant to the set $S$. Initially, the clustering consists of singleton sets:

$$\pi^1 = \{\{\mathbf{o}_i\} \mid 1 \leq i \leq 7\}\{\{\mathbf{o}_1\}, \{\mathbf{o}_2\}, \{\mathbf{o}_3\}, \{\mathbf{o}_4\}, \{\mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}\}.$$

Two of the closest clusters are $\{\mathbf{o}_1\}, \{\mathbf{o}_2\}$; these clusters are fused into the cluster $\{\mathbf{o}_1, \mathbf{o}_2\}$, the new partition is

$$\pi^2 = \{\{\mathbf{o}_1, \mathbf{o}_2\}, \ldots, \{\mathbf{o}_7\}\},$$

and the matrix of dissimilarities becomes the $6 \times 6$-matrix

$$D^2 = \begin{pmatrix} 0 & 2 & 5 & 7 & 10 & 9 \\ 2 & 0 & 3 & 5 & 8 & 7 \\ 5 & 3 & 0 & 2 & 5 & 4 \\ 7 & 5 & 2 & 0 & 3 & 4 \\ 10 & 8 & 5 & 3 & 0 & 3 \\ 9 & 7 & 4 & 4 & 3 & 0 \end{pmatrix}.$$

The next pair of closest clusters is $\{\mathbf{o}_1, \mathbf{o}_2\}$ and $\{\mathbf{o}_3\}$. These clusters are fused into the cluster $\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}$, and the new $5 \times 5$-matrix is:

$$D^3 = \begin{pmatrix} 0 & 3 & 5 & 8 & 7 \\ 3 & 0 & 2 & 5 & 4 \\ 5 & 2 & 0 & 3 & 4 \\ 8 & 5 & 3 & 0 & 3 \\ 7 & 4 & 4 & 3 & 0 \end{pmatrix},$$

which corresponds to the partition

$$\pi^3 = \{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4\}, \ldots, \{\mathbf{o}_7\}\}.$$

Next, the closest clusters are $\{\mathbf{o}_4\}$ and $\{\mathbf{o}_5\}$. Fusing these yields the partition

$$\pi^4 = \{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}\}$$

and the $4 \times 4$-matrix

$$D^4 = \begin{pmatrix} 0 & 3 & 8 & 7 \\ 3 & 0 & 3 & 4 \\ 8 & 3 & 0 & 3 \\ 7 & 4 & 3 & 0 \end{pmatrix}$$

We have three choices now since there are three pairs of clusters at distance 3 of each other: $(\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\})$, $(\{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6\})$, and $(\{\mathbf{o}_6\}, \{\mathbf{o}_7)$. By choosing to fuse the first pair, we obtain the partition

$$\pi^5 = \{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}\},$$

which corresponds to the $3 \times 3$-matrix

$$D^5 = \begin{pmatrix} 0 & 3 & 4 \\ 3 & 0 & 3 \\ 4 & 3 & 0 \end{pmatrix}.$$

Observe that the large cluster $\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4, \mathbf{o}_5\}$ formed so far has an elongated shape, which is typical for single-link variant of the algorithm.

Next, we coalesce $\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4, \mathbf{o}_5\}$ and $\{\mathbf{o}_6\}$, which yields

$$\pi^6 = \{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4, \mathbf{o}_5, \mathbf{o}_6\}, \{\mathbf{o}_7\}\}$$

and

$$D^6 = \begin{pmatrix} 0 & 3 \\ 3 & 0 \end{pmatrix}.$$

Finally, we join the last two clusters, and the clustering is completed.

The dendrogram of the hierarchy produced by the algorithm is given in Figure 13.2.

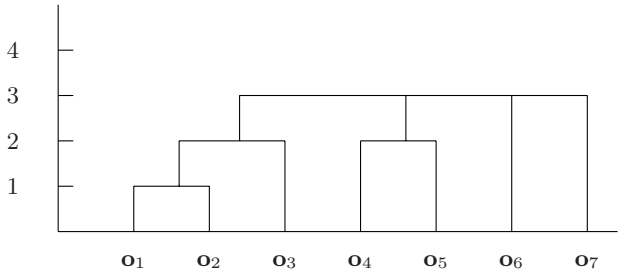



**Fig. 13.2.** Dendrogram of single-link clustering.

The variant of the algorithm that uses the function *cl* is known as the *complete-link* clustering. It tends to favor globular clusters.

*Example 13.5.* Now we apply the complete-link algorithm to the set $S$ considered in Example 13.4. It is easy to see that the initial two partitions and the initial matrix are the same as for the single-link algorithm.

However, after creating the first cluster $\{\mathbf{o}_1, \mathbf{o}_2\}$, the distance matrices begin to differ. The next matrix is

$$D^2 = \begin{pmatrix} 0 & 3 & 6 & 8 & 11 & 10 \\ 3 & 0 & 3 & 5 & 8 & 7 \\ 6 & 3 & 0 & 2 & 5 & 4 \\ 8 & 5 & 2 & 0 & 3 & 4 \\ 11 & 8 & 5 & 3 & 0 & 3 \\ 10 & 7 & 4 & 4 & 3 & 0 \end{pmatrix},$$

which shows that the closest clusters are now $\{\mathbf{o}_4\}$ and $\{\mathbf{o}_5\}$. Thus,

$$\pi^3 = \{\{\mathbf{o}_1, \mathbf{o}_2\}, \{\mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}\}$$

and the new matrix is

$$D^3 = \begin{pmatrix} 0 & 3 & 8 & 11 & 10 \\ 3 & 0 & 5 & 8 & 7 \\ 8 & 5 & 0 & 5 & 3 \\ 11 & 8 & 5 & 0 & 3 \\ 10 & 7 & 3 & 3 & 0 \end{pmatrix}.$$

Three pairs of clusters correspond to the minimal value 3 in $D^3$:

$$(\{\mathbf{o}_1, \mathbf{o}_2\}, \{\mathbf{o}_3\}),$$
$$(\{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_3\}),$$
$$(\{\mathbf{o}_6\}, \{\mathbf{o}_7\}).$$

If we merge the last pair, we get the partition

$$\pi^4 = \{\{\mathbf{o}_1, \mathbf{o}_2\}, \{\mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6, \mathbf{o}_7\}\}$$

and the matrix

$$D^4 = \begin{pmatrix} 0 & 3 & 8 & 11 \\ 3 & 0 & 5 & 8 \\ 8 & 5 & 0 & 5 \\ 11 & 8 & 5 & 0 \end{pmatrix}.$$

Next, the closest clusters are $\{\mathbf{o}_1, \mathbf{o}_2\}, \{\mathbf{o}_3\}$. Merging those clusters will result in the partition $\pi^5 = \{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6, \mathbf{o}_7\}\}$ and the matrix

$$D^5 = \begin{pmatrix} 0 & 8 & 11 \\ 8 & 0 & 5 \\ 11 & 5 & 0 \end{pmatrix}.$$

The current clustering is shown in Figure 13.3. Observe that in the case of the clusters obtained by the complete-link method that appear early tend to enclose objects that are closed in the sense of the distance.

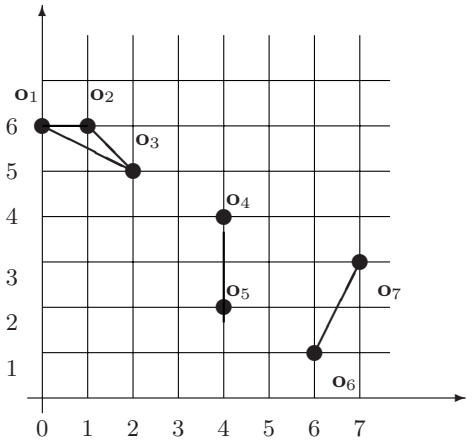



**Fig. 13.3.** Partial clustering obtained by the complete-link method.

Now the closest clusters are $\{\mathbf{o}_4, \mathbf{o}_5\}$ and $\{\mathbf{o}_6, \mathbf{o}_7\}$. By merging those clusters, we obtain the partition $\pi^5 = \{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5, \mathbf{o}_6, \mathbf{o}_7\}\}$ and the matrix

$$D^6 = \begin{pmatrix} 0 & 11 \\ 11 & 0 \end{pmatrix}.$$

The dendrogram of the resulting clustering is given in Figure 13.4.

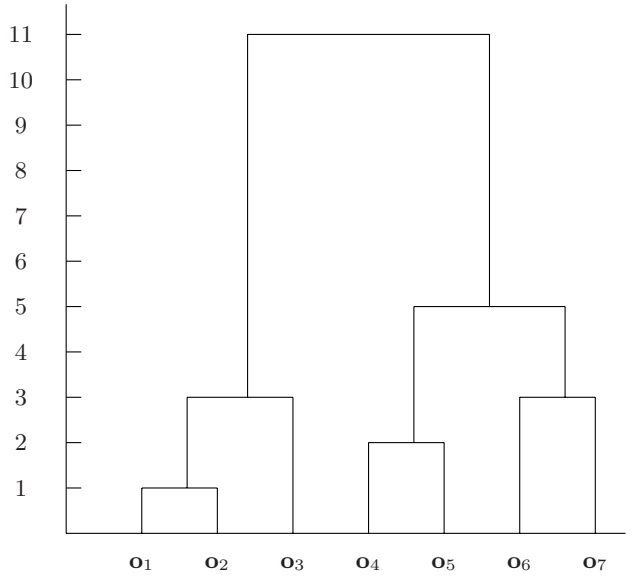



**Fig. 13.4.** Dendrogram of complete-link clustering.

The *group average method*, which makes use of the *gav* function generates an intermediate approach between the single-link and the complete-link method. What the methods mentioned so far have in common is the *monotonicity property* expressed by the following statement.

**Theorem 13.6.** *Let $(S, d)$ be a finite metric space and let $D^1, \ldots, D^m$ be the sequence of matrices constructed by any of the first three hierarchical methods (single, complete, or average link), where $m = |S|$. If $\mu_i$ is the smallest entry of the matrix $D^i$ for $1 \leq i \leq m$, then $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_m$. In other words, the dissimilarity between clusters that are merged at each step is nondecreasing.*

*Proof.* Suppose that the matrix $D^{j+1}$ is obtained from the matrix $D^j$ by merging the clusters $C_p$ and $C_q$ that correspond to the lines $p$ and $q$ and to columns $p, q$ of $D^j$. This happens because $d_{pq} = d_{qp}$ is one of the minimal elements of the matrix $D^j$. Then, these lines and columns are replaced with a line and column that corresponds to the new cluster $C_r$ and the dissimilarities

between this new cluster and the previous clusters $C_i$, where $i \neq p, q$. The elements $d_{rh}^{j+1}$ of the new line (and column) are obtained either as $\min\{d_{ph}^j, d_{qh}^j\}$, $\max\{d_{ph}^j, d_{qh}^j\}$, or $\frac{|C_p|}{|C_r|}d_{ph}^j + \frac{|C_q|}{|C_r|}d_{qh}^j$, for the single-link, complete-link, or group average methods, respectively. In any of these cases, it is not possible to obtain a value for $d_{rh}^{j+1}$ that is less than the minimal value of an element of $D^j$. □

The last two methods captured by the Lance-Williams formula are the centroid method and the Ward method of clustering. As we observed before, Formula (13.2) shows that the dissimilarity of two clusters in the case of Ward's method equals the increase in the sum of the squared errors that results when the clusters are merged. The centroid method adopts the distance between the centroids as the distance between the corresponding clusters. Either method lacks the monotonicity properties.

### 13.2.2 Graph-based Hierarchical Clustering

Starting from a finite metric space $(S, d)$, we can construct a sequence of *threshold graphs* $\mathcal{G}_0, \ldots, \mathcal{G}_k$, where $k = diam_{S,d}$ is the diameter of $(S, d)$. The threshold graph $\mathcal{G}_p = (S, E_p)$ is defined by its set of edges

$$E_p = \{(x, y) \in S \times S \mid d(x, y) \leq p\}$$

for $0 \leq p \leq k$. The graph $\mathcal{G}_0$ is $(S, \emptyset)$, while $\mathcal{G}_k$ is a complete graph on the set $S$.

*Example 13.7.* The sequence of threshold graphs for the metric space $(S, d_1)$ introduced in Example 13.4 is given in Figures 13.5 and 13.6.

A variant of single-link clustering can now be achieved by working on the sequence of threshold graphs. Let $\mathcal{G}_0, \ldots, \mathcal{G}_k$ be the sequence of threshold graphs of a finite metric space, where $k = diam_{S,d}$, and let $c_i$ be the number of connected components of the threshold graph $\mathcal{G}_i$ for $1 \leq i \leq k$.

**Algorithm 13.8 (Graph-Based Single-Link Clustering)**
**Input:** a finite metric space $(S, d)$ of diameter $k$, where $|S| = n$.
**Output:** a hierarchy of clusters on $S$.
**Method:**
initialize the threshold graph $\mathcal{G}_0$;
$c = n$; // current number of connected components
$p = 1$;
**while** $(c_p > 1)$ do
  **if** $(c_p < c)$ then
    output the connected components $\mathcal{G}_p$;
  $p$++;
**endwhile**

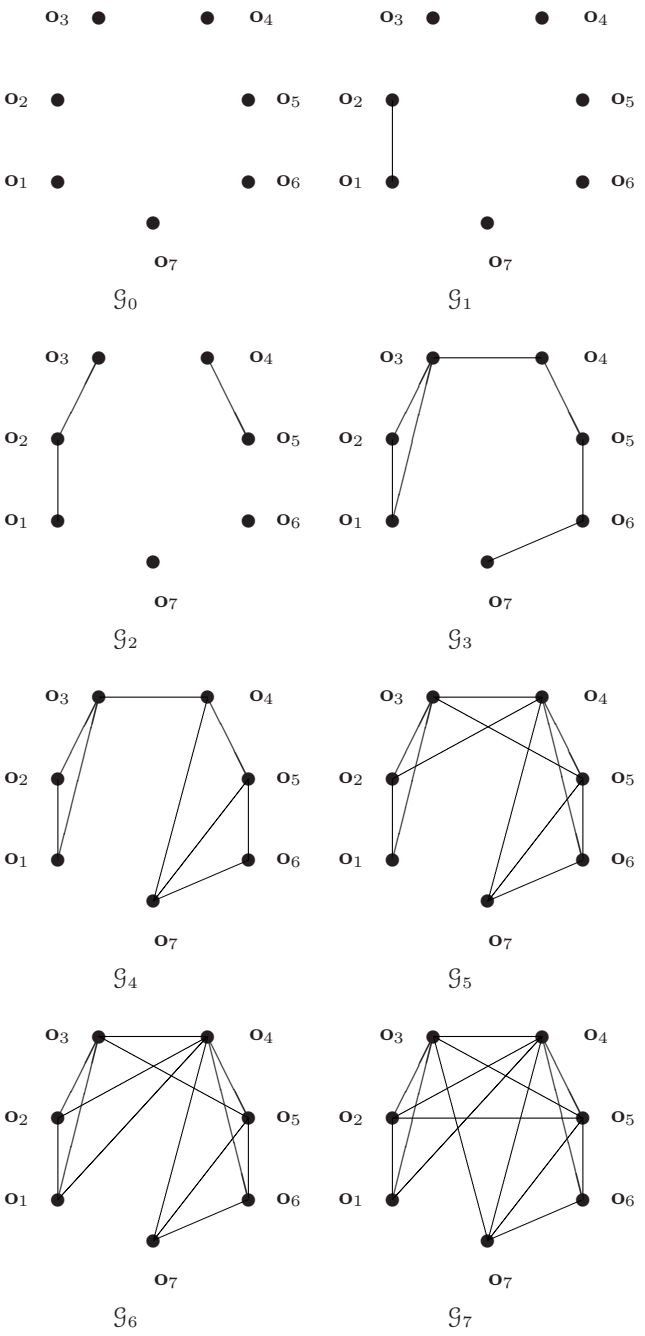



**Fig. 13.5.** Threshold graphs of $(S, d)$.

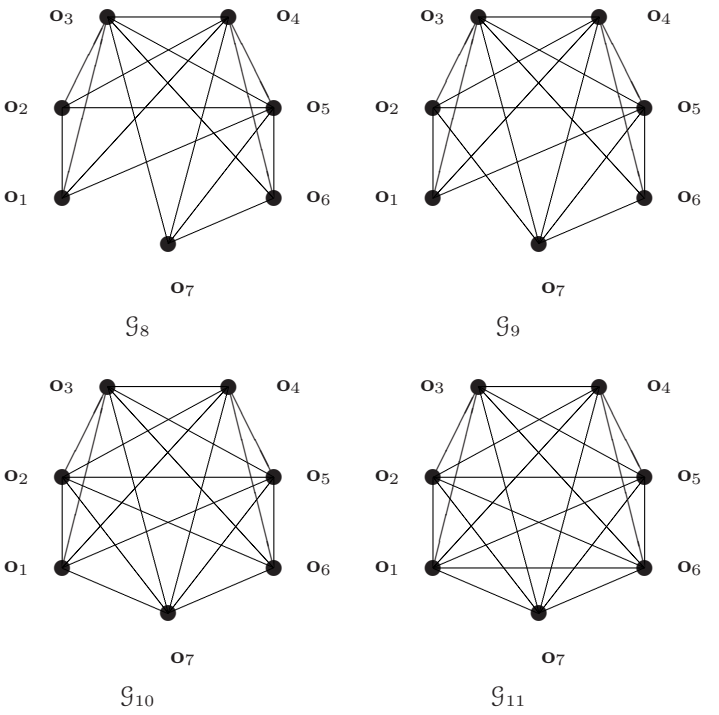



**Fig. 13.6.** Threshold graphs of $(S, d)$ (continued).

Observe that in Algorithm 13.8 some of the connected components of a threshold graph $\mathcal{G}_{p-1}$ persist as connected components of the graph $\mathcal{G}_p$. Others will coalesce to form larger clusters, and this is possible if at least one edge of weight $p$ exists between the clusters that are about to combine. This explains the term "single-link" used to designate this algorithm (see [56]).

*Example 13.9.* For the threshold graphs of Example 13.7, the connected components are shown in the next table:

| $i$ | Connected Components of $\mathcal{G}_i$ | $c_i$ |
|---|---|---|
| 0 | $\{\mathbf{o}_1\}, \{\mathbf{o}_2\}, \{\mathbf{o}_3\}, \{\mathbf{o}_4\}, \{\mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}$ | 7 |
| 1 | $\{\mathbf{o}_1, \mathbf{o}_2\}, \{\mathbf{o}_3\}, \{\mathbf{o}_4\}, \{\mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}$ | 6 |
| 2 | $\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}$ | 4 |
| 3 | $\{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_7\}$ | 1 |

**Algorithm 13.10 (Graph-Based Complete-link Clustering)**
**Input:** a finite metric space $(S, d)$ of diameter $k$, where $|S| = n$;

**Output:** a hierarchy of clusters on $S$;
**Method:**
initialize the threshold graph $\mathcal{G}_0$;
initialize the collection of clusters to all singleton sets;
**while** $(p < k)$ do
  **for all pairs of disjoint clusters** $(P, Q)$ in $\mathcal{G}_p$ **do**
    **if** $P \cup Q$ is a clique in $\mathcal{G}_{p+1}$ then
      add $P \cup Q$ to the set of clusters;
  $p$++;
**endwhile**

Note that not all cliques that form in the threshold graphs are produced as clusters by Algorithm 13.10. Indeed, the algorithm generates a cluster only when two disjoint clusters that have already appeared are cliques and will form a new clique by the addition of edges that have been added in successive threshold graphs. In this manner, the algorithm indeed produces a hierarchy of clusters.

*Example 13.11.* In Table 13.1, we show the cluster formation in the graph-based clustering algorithm. Observe that the cluster $\{\mathbf{o}_4, \mathbf{o}_5, \mathbf{o}_6, \mathbf{o}_7\}$ is produced at $p = 5$, when in intermediate threshold graphs $\mathcal{G}_3$ and $\mathcal{G}_4$ sufficient edges were added to allow us to join the disjoint clusters $\{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6, \mathbf{o}_7\}$ into the new cluster.

**Table 13.1.** Clusters produced by the complete-link graph-based algorithm

| $p$ | Clusters in $\mathcal{G}_p$ |
|---|---|
| 0 | $\{\{\mathbf{o}_1\}, \{\mathbf{o}_2\}, \{\mathbf{o}_3\}, \{\mathbf{o}_4\}, \{\mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}\}$ |
| 1 | $\{\{\mathbf{o}_1, \mathbf{o}_2\}, \{\mathbf{o}_3\}, \{\mathbf{o}_4\}, \{\mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}\}$ |
| 2 | $\{\{\mathbf{o}_1, \mathbf{o}_2\}, \{\mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}\}$ |
| 3 | $\{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6, \mathbf{o}_7\}\}$ |
| 4 | $\{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6, \mathbf{o}_7\}\}$ |
| 5 | $\{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5, \mathbf{o}_6, \mathbf{o}_7\}\}$ |
| ⋮ | ⋮ |
| 10 | $\{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5, \mathbf{o}_6, \mathbf{o}_7\}\}$ |
| 11 | $\{\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4, \mathbf{o}_5, \mathbf{o}_6, \mathbf{o}_7\}\}$ |

The usefulness of the application of minimal spanning trees to clustering was explored by C. T. Zahn [147].

Let $S = \{o_1, \ldots, o_n\}$ be a set of $n$ objects and let $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ be a metric. The weighted complete graph $(\mathcal{K}_n, w)$ is defined by $\mathcal{K}_n = (S, \{(o_i, o_j) \mid i \neq j\})$, and $w(o_i, o_j) = d(o_i, o_j)$ for $1 \leq i, j \leq n$ and $i \neq j$.

The notion of an *inconsistent edge* is essential for Zahn's algorithm and there are several plausible definitions for it. For instance, we can define an

edge to be inconsistent if its weight is larger by a certain factor than the averages of the weights of both sides of $(x, y)$.

**Algorithm 13.12 (Zahn's Clustering Algorithm)**
**Input:** a complete graph $(\mathcal{K}_n, w)$, where
$\mathcal{K}_n = (S, \{(o_i, o_j) \mid i \neq j\})$, and
$w(o_i, o_j) = d(o_i, o_j)$ for $1 \leq i, j \leq n$ and $i \neq j$;
**Output:** a clustering of the objects of $S$.
**Method:**
construct a minimal spanning tree for $(\mathcal{K}_n, w)$;
identify inconsistent edges in the minimal spanning tree;
create a cluster hierarchy by successively removing inconsistent edges.

Zahn's algorithm leads to a hierarchy on the set of vertices since each removal of an edge from a tree creates two disjoint connected components of that tree, that are themselves trees.

*Example 13.13.* Again, we are using the set of objects introduced in Example 13.4 equipped with the $d_1$ metric given by the weight matrix

$$D^1 = \begin{pmatrix} 0 & 1 & 3 & 6 & 8 & 11 & 10 \\ 1 & 0 & 2 & 5 & 7 & 10 & 9 \\ 3 & 2 & 0 & 3 & 5 & 8 & 7 \\ 6 & 5 & 3 & 0 & 2 & 5 & 4 \\ 8 & 7 & 5 & 2 & 0 & 3 & 4 \\ 11 & 10 & 8 & 5 & 3 & 0 & 3 \\ 10 & 9 & 7 & 4 & 4 & 3 & 0 \end{pmatrix}.$$

Kruskal's algorithm (Algorithm 3.48) applied to the weighted complete graph $(\mathcal{K}_7, w)$ yields the minimal spanning tree shown in Figure 13.7. Suppose that an edge is deemed to be inconsistent if its weight is larger than the average of the adjacent edges. Then the most inconsistent edge is $(\mathbf{o}_3, \mathbf{o}_4)$ of weight 3, which is 1.5 times the average of the adjacent edges. By removing this edge the tree is divided into two connected components, $\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \}$ and $\{\mathbf{o}_4, \mathbf{o}_5, \mathbf{o}_6, \mathbf{o}_7\}$. Among the edges of the first component, the edge $(\mathbf{o}_2, \mathbf{o}_3)$ is the most inconsistent. By removing it, we get the clusters $\{\mathbf{o}_1, \mathbf{o}_2\}$ and $\{\mathbf{o}_3\}$. Similarly, by removing the edge $(\mathbf{o}_5, \mathbf{o}_6)$, the second cluster is split into the clusters $\{bfo_4, \mathbf{o}_5\}$ and $\{bfo_6, \mathbf{o}_7\}$.

There exists an interesting link between the single-link clustering algorithm and the subdominant ultrametric of a dissimilarity, which we examined in Section 10.4.

To construct the subdominant ultrametric for a dissimilarity space $(S, d)$, we built an increasing chain of partitions $\pi_1, \pi_2, \ldots$ of $S$ (where $\pi_1 = \alpha_S$)

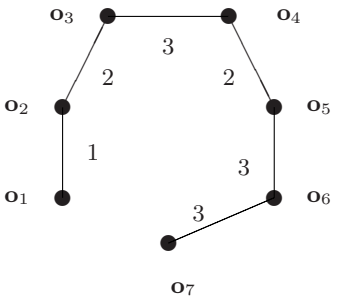



**Fig. 13.7.** Spanning tree of the graph $(\mathcal{K}_7, w)$.

and a sequence of dissimilarities $d_1, d_2, \ldots$ (where $d_1 = d$) on the sets of blocks of $\pi_1, \pi_2, \ldots$, respectively. We claim that this sequence of partitions $\pi_1, \pi_2, \ldots$ coincides with the sequence of partitions $\pi^1, \pi^2, \ldots$, and that the sequence of dissimilarities $d_1, d_2, \ldots$ coincides with the sequences of dissimilarities $d^1, d^2, \ldots$ defined by the matrices $D^i$ constructed by the single-link algorithm. This is clearly the case for $i = 1$.

Suppose that the statement is true for $i$. The partition $\pi_{i+1}$ is obtained from $\pi_i$ by fusing the blocks $B, C$ of $\pi$ such that $d_i(B, C)$ has the smallest value, that is,

$$\pi_{i+1} = (\pi_i - \{B, C\}) \cup \{B \cup C\}.$$

Since this is exactly how the partition $\pi^{i+1}$ is constructed from $\pi^i$, it follows that $\pi_{i+1} = \pi^{i+1}$. The inductive hypothesis implies that

$$d^i(U, V) = d_i(U, V) = \min\{d(u, v) \mid u \in U, v \in V\}$$

for all $U, V \in \pi_i$. Since the dissimilarity $d_{i+1}$ is $d_{i+1}(U, V) = \min\{d(u, v) \mid u \in U, u \in V\}$ for every pair of blocks $U, V$ of $\pi_{i+1}$, it is clear that $d_{i+1}(U, V) = d_i(U, V) = d^i(U, V) = d^{i+1}(U, V)$ when neither $U$ nor $V$ equal the block $B \cup C$. Then,

$$\begin{aligned}
&d_{i+1}(B \cup C, W)\\
&= \min\{d(t, w) \mid t \in B \cup C, w \in W\}\\
&= \min\{\min\{d(b, w) \mid b \in B, w \in W\}, \min\{d(c, w) \mid c \in C, w \in W\}\}\\
&= \min\{d_i(B, W), d_i(C, W)\}\\
&= \min\{d^i(B, W), d^i(C, W)\}\\
&= d^{i+1}(B \cup C, W).
\end{aligned}$$

Thus, $d_{i+1} = d^{i+1}$.

Let $x, y$ be a pair of elements of $S$. The value of the subdominant ultrametric is given by

$$e(x, y) = \min\{h_d(W) \mid W \in \mathcal{H}_d \text{ and } \{x, y\} \subseteq W\}.$$

This is the height of $W$ in the dendrogram of the single-link clustering and therefore the subdominant ultrametric can be read directly from this dendrogram.

*Example 13.14.* The subdominant ultrametric of the Euclidean metric considered in Example 13.4 is given by the following table:

| $e(\mathbf{o}_i, \mathbf{o}_j)$ | $\mathbf{o}_1$ | $\mathbf{o}_2$ | $\mathbf{o}_3$ | $\mathbf{o}_4$ | $\mathbf{o}_5$ | $\mathbf{o}_6$ | $\mathbf{o}_7$ |
|---|---|---|---|---|---|---|---|
| $\mathbf{o}_1$ | 0 | 1 | 2 | 3 | 3 | 3 | 3 |
| $\mathbf{o}_2$ | 1 | 0 | 2 | 3 | 3 | 3 | 3 |
| $\mathbf{o}_3$ | 2 | 2 | 0 | 3 | 3 | 3 | 3 |
| $\mathbf{o}_4$ | 3 | 3 | 3 | 0 | 2 | 3 | 3 |
| $\mathbf{o}_5$ | 3 | 3 | 3 | 2 | 0 | 3 | 3 |
| $\mathbf{o}_6$ | 3 | 3 | 3 | 3 | 3 | 0 | 3 |
| $\mathbf{o}_7$ | 3 | 3 | 3 | 3 | 3 | 3 | 0 |

## 13.3 The *k*-Means Algorithm

The $k$-means algorithm is a partitional algorithm that requires the specification of the number of clusters $k$ as an input. The set of objects to be clustered $S = \{\mathbf{o}^1, \ldots, \mathbf{o}^n\}$ is a subset of $\mathbb{R}^m$. Due to its simplicity and its many implementations it is a very popular algorithm despite this requirement.

The $k$-means algorithm begins with a randomly chosen collection of $k$ points $\mathbf{c}^1, \ldots, \mathbf{c}^k$ in $\mathbb{R}^m$ called centroids. An initial partition of the set $S$ of objects is computed by assigning each object $\mathbf{o}^i$ to its closest centroid $\mathbf{c}^j$. Let $U_j$ be the set of points assigned to the centroid $\mathbf{c}^j$.

The assignments of objects to centroids are expressed by a matrix $(b_{ij})$, where

$$b_{ij} = \begin{cases} 1 & \text{if } \mathbf{o}^i \in U_j, \\ 0 & \text{otherwise.} \end{cases}$$

Since each object is assigned to exactly one cluster, we have $\sum_{j=1}^{k} b_{ij} = 1$. On the other hand, $\sum_{i=1}^{n} b_{ij}$ equals the number of objects assigned to the centroid $\mathbf{c}^j$.

After these assignments, expressed by the matrix $(b_{ij})$, the centroids $\mathbf{c}^j$ must be re-computed using the formula:

$$\mathbf{c}^j = \frac{\sum_{i=1}^{n} b_{ij}\mathbf{o}^i}{\sum_{i=1}^{n} b_{ij}} \tag{13.3}$$

for $1 \leq j \leq k$.

The sum of squared errors of a partition $\pi = \{U_1, \ldots, U_k\}$ of a set of objects $S$ was defined in Equality (13.1) as

$$sse(\pi) = \sum_{j=1}^{k} \sum_{\mathbf{o} \in U_j} d^2(\mathbf{o}, \mathbf{c}^j),$$

where $\mathbf{c}^j$ is the centroid of $U_j$ for $1 \leq j \leq k$. The error of such an assignment is the sum of squared errors of the partition $\pi = \{U_1, \ldots, U_k\}$ defined as

$$\begin{aligned} sse(\pi) &= \sum_{i=1}^{n} \sum_{j=1}^{k} b_{ij} ||\mathbf{o}^i - \mathbf{c}^j||^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} b_{ij} \sum_{p=1}^{m} \left(o_p^i - c_p^j\right)^2. \end{aligned}$$

The $mk$ necessary conditions for a local minimum of this function,

$$\frac{\partial sse(\pi)}{\partial c_p^j} = \sum_{i=1}^{n} b_{ij} \left(-2(o_p^i - c_p^j)\right) = 0,$$

for $1 \leq p \leq m$ and $1 \leq j \leq k$, can be written as

$$\sum_{i=1}^{n} b_{ij} o_p^i = \sum_{i=1}^{n} b_{ij} c_p^j = c_p^j \sum_{i=1}^{n} b_{ij},$$

or as

$$c_p^j = \frac{\sum_{i=1}^{n} b_{ij} o_p^i}{\sum_{i=1}^{n} b_{ij}}$$

for $1 \leq p \leq m$. In vectorial form, these conditions amount to

$$\mathbf{c}^j = \frac{\sum_{i=1}^{n} b_{ij} \mathbf{o}^i}{\sum_{i=1}^{n} b_{ij}},$$

which is exactly the formula (13.3) that is used to update the centroids. Thus, the choice of the centroids can be justified by the goal of obtaining local minima of the sum of squared errors of the clusterings.

Since we have new centroids, objects must be reassigned, which means that the values of $b_{ij}$ must be recomputed, which in turn will affect the values of the centroids, etc.

The halting criterion of the algorithm depends on particular implementations and may involve

(i) performing a certain number of iterations;
(ii) lowering the sum of squared errors $sse(\pi)$ below a certain limit;
(iii) the current partition coinciding with the previous partition.

This variant of the $k$-means algorithm is known as *Forgy's* algorithm.

**Algorithm 13.15 (Forgy's Algorithm)**
**Input:** set of objects to be clustered $S = \{\mathbf{o}^1, \ldots, \mathbf{o}^n\}$ and the number of clusters $k$;
**Output:** a collection of $k$ clusters;
**Method:**
obtain a randomly chosen collection of $k$ points $\mathbf{c}_1, \ldots, \mathbf{c}_k$ in $\mathbb{R}^n$;
assign each object $\mathbf{o}^i$ to the closest centroid $\mathbf{c}^j$;
let $\pi = \{U_1, \ldots, U_k\}$ be the partition defined by $\mathbf{c}^1, \ldots, \mathbf{c}^k$;
recompute the centroids of the clusters $U_1, \ldots, U_k$;
**while** (halting criterion is not met) **do**
compute the new value of the partition $\pi$
using the current centroids;
recompute the centroids of the blocks of $\pi$;
**endwhile**

The popularity of the $k$-means algorithm stems from its simplicity and its low time complexity $O(kn\ell)$, where $n$ is the number of objects to be clustered and $\ell$ is the number of iterations that the algorithm is performing.

Another variant of the $k$-means algorithm redistributes objects to clusters based on the effect of such a reassignment on the objective function. If $sse(\pi)$ decreases, the object is moved and the two centroids of the affected clusters are recomputed. This variant is carefully analyzed in [12]

## 13.4 The PAM Algorithm

Another algorithm, named PAM (an acronym of "Partition Around Medoids") developed by Kaufman and Rousseeuw [76], also requires as an input parameter the number $k$ of clusters to be extracted.

The $k$ clusters are determined based on a representative object from each cluster, called the *medoid* of the cluster. The medoid is intended to have the most central position in the cluster relative to all other members of the cluster. Once medoids are selected, each remaining object $o$ is assigned to a cluster represented by a medoid $o_i$ if the dissimilarity $d(o, o_i)$ is minimal.

In a second phase, swapping objects and existing medoids is considered. A cost of a swap is defined with the intention of penalizing swaps that diminish the centrality of the medoids in the clusters. Swapping continues as long as useful swaps (that is, swaps with negative costs) can be found.

PAM begins with a set of objects $S$, where $|S| = n$, a dissimilarity $n \times n$ matrix $D$, and a prescribed number of clusters $k$. The $d_{ij}$ entry of the matrix $D$ is the dissimilarity $d(o_i, o_j)$ between the objects $o_i$ and $o_j$. PAM is more robust than Forgy's variant of $k$-clustering because it minimizes the sum of the dissimilarities instead of the sum of the squared errors.

The algorithm has two distinct phases: the *building phase* and the *swapping phase*. The building phase aims to construct a set $L$ of selected objects, $L \subseteq S$.

The set of remaining objects is denoted by $R$; clearly, $R = S - L$. We begin by determining the most centrally located object.

The quantities $Q_i = \sum_{j=1}^{n} d_{ij}$ are computed starting from the matrix $D$. The most central object $o_q$ is determined by

$$q = \arg\min_i Q_i.$$

The set $L$ is initialized as $L = \{o_q\}$.

Suppose now that we have constructed a set $L$ of selected objects and $|L| < k$. We need to add a new selected object to the set $L$. To do this, we need to examine all objects that have not been included in $L$ so far, that is, all objects in $R$. The selection is determined by a merit function $M : R \longrightarrow \mathbb{N}$.

To compute the merit $M(o)$ of an object $o \in R$, we scan all objects in $R$ distinct from $o$. Let $o' \in R - \{o\}$ be such an object. If $d(o, o') < d(L, o')$, then adding $o$ to $L$ could benefit the clustering (from the point of view of $o'$) because $d(L, o')$ will diminish. The potential benefit is $d(o', L) - d(o, o')$. Of course, if $d(o, o') \geq d(L, o')$, no such benefit exists (from the point of view of $o'$). Thus, we compute the merit of $o$ as

$$M(o) = \sum_{o' \in R - \{o\}} \max\{D(L, o') - d(o, o'), 0\}.$$

We add to $L$ the unselected object $o$ that has the largest merit value. The building phase halts when $|L| = k$.

The objects in set $L$ are the potential medoids of the $k$ clusters that we seek to build. The second phase of the algorithm aims to improve the clustering by considering the merit of swaps between selected and unselected objects. So, assume now that $o_i$ is a selected object, $o_i \in L$, and $o_h$ is an unselected object, $o_h \in R = S - L$. We need to determine the cost $C(o_i, o_h)$ of swapping $o_i$ and $o_h$. Let $o_j$ be an arbitrary unselected object. The contribution $c_{ihj}$ of $o_j$ to the cost of the swap between $o_i$ and $o_h$ is defined as follows:

1. If $d(o_i, o_j)$ and $d(o_h, o_j)$ are greater than $d(o, o_j)$ for any $o \in L - \{o_i\}$, then $c_{ihj} = 0$.
2. If $d(o_i, o_j) = d(L, o_j)$, then two cases must be considered depending on the distance $e(o_j)$ from $e_j$ to the second-closest object of $S$.
   a) If $d(o_h, o_j) < e(o_j)$, then $c_{ihj} = d(o_h, o_j) - d(S, o_j)$.
   b) If $d(o_h, o_j) \geq e(o_j)$, then $c_{ihj} = e(o_j) - d(S, o_j)$.
   In either of these two subcases, we have
   $$c_{ihj} = \min\{d(o_h, o_j), e_j\} - d(o_i, o_j).$$
3. If $d(o_i, o_j) > d(L, o_j)$ (that is, $o_j$ is more distant from $o_i$ than from at least one other selected object) and $d(o_h, o_j) < d(L, o_j)$ (which means that $o_j$ is closer to $o_h$ than to any selected object), then $c_{ihj} = d(o_h, o_j) - d(S, o_j)$.

The cost of the swap is $C(o_i, o_h) = \sum_{o_j \in R} c_{ihj}$. The pair that minimizes $C(o_i, o_j)$ is selected. If $C(o_i, o_j) < 0$, then the swap is carried out. All potential swaps are considered.

The algorithm halts when no useful swap exists; that is, no swap with negative cost can be found.

The pseudocode of the algorithm follows.

**Algorithm 13.16 k_means_PAM**{
construct the set $L$ of $k$ medoids;
**repeat**
compute the costs $C(o_i, o_h)$ for $o_i \in L$ and $o_h \in R$;
select the pair $(o_i, o_h)$ that corresponds to the minimum $m = C(o_i, o_h)$;
**until** $(m > 0)$;
}

Note that inside the loop **repeat** $\cdots$ **until** there are $l(n-l)$ pairs of objects to be examined, and for each pair we need to involve $n-l$ non-selected objects. Thus, one execution of the loop requires $O(l(n-l)^2)$, and the total execution may require up to $O\left(\sum_{l=1}^{n-l} l(n-l)^2\right)$, which is $O(n^4)$. Thus, the usefulness of PAM is limited to rather small data set (no more than a few hundred objects).

## 13.5 Limitations of Clustering

As we stated before, an exclusive clustering of a set of objects $S$ is a partition of $S$ whose blocks are the *clusters*. A clustering method starts with a definite dissimilarity on $S$ and generates a clustering. This is formalized in the next definition.

**Definition 13.17.** *Let $S$ be a set of objects and let $\mathcal{D}'_S$ be the set of definite dissimilarities that can be defined on $S$.*

*A clustering function on $S$ is a mapping $f : \mathcal{D}'_S \longrightarrow PART(S)$.*

*Example 13.18.* Let $g : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ be a continuous, nondecreasing and unbounded function and let $S \subseteq \mathbb{R}^n$ be a finite subset of $\mathbb{R}^n$. For $k \in \mathbb{N}$ and $k \geq 2$, define a $(g, k)$-clustering function as follows.

Begin by selecting a set $T$ of $k$ points from $S$ such that the function $\Lambda_d^g(T) = \sum_{x \in S} g(d(x, T))$ is minimized. Here $d(x, T) = \min\{d(x, t) | t \in T\}$. Then, define a partition of $S$ into $k$ clusters by assigning each point to the point in $T$ that is the closest and breaking the ties using a fixed (but otherwise arbitrary) order on the set of points. The clustering function defined by $(d, g)$, denoted by $f^g$, maps $d$ to this partition.

The *k-median clustering function*, is obtained by choosing $g(x) = x$ for $x \in \mathbb{R}_{\geq 0}$; the *k-means clustering function* is obtained by taking $g(x) = x^2$ for $x \in \mathbb{R}_{\geq 0}$.

**Definition 13.19.** *Let $\kappa$ be a partition of $S$ and let $d, d' \in \mathcal{D}'_S$. The definite dissimilarity $d'$ is a $\kappa$-transformation of $d$ if the following conditions are satisfied:*
(i) *if $x \equiv_\kappa y$, then $d'(x, y) \leq d(x, y)$;*
(ii) *if $x \not\equiv_\kappa y$, then $d'(x, y) > d(x, y)$.*

In other words, $d'$ is a $\kappa$-transformation of $d$ if for two objects that belong to the same $\kappa$-cluster $d'(x, y)$ is smaller than $d(x, y)$, while for two objects that belong to two distinct clusters $d'(x, y)$ is larger than $d(x, y)$.

Next, we consider three desirable properties of a clustering function.

**Definition 13.20.** *Let $S$ be a set and let $f : \mathcal{D}'_S \longrightarrow PART(S)$ be a clustering function. The function $f$ is*
(i) *scale-invariant if, for every $d \in \mathcal{D}'_S$ and every $\alpha > 0$, we have $f(d) = f(\alpha d)$;*
(ii) *rich, if $\mathrm{Ran}(f) = PART(S)$;*
(iii) *consistent if, for every $d, d' \in \mathcal{D}'_S$ and $\kappa \in PART(S)$ such that $f(d) = \kappa$ and $d'$ is a $\kappa$-transformation of $d$, we have $f(d') = \kappa$,*

Unfortunately, as we shall see in Theorem 13.25, established in [80], there is no clustering function that enjoys all three properties.

The following definition will be used in the proof of Lemma 13.23.

**Definition 13.21.** *A dissimilarity $d \in \mathcal{D}'_S$ is $(a, b)$-conformant to a clustering $\kappa$ if $x \equiv_\kappa y$ implies $d(x, y) \leq a$ and $x \not\equiv_\kappa y$ implies $d(x, y) \geq b$.*

*A dissimilarity is conformant to a clustering $\kappa$ if it is $(a, b)$-conformant to $\kappa$ for some pair of numbers $(a, b)$.*

Note that if $d'$ is a $\kappa$-transformation of $d$, and $d$ is $(a, b)$-conformant to $\kappa$, then $d'$ is also $(a, b)$-conformant to $\kappa$.

**Definition 13.22.** *Let $\kappa \in PART(S)$ be a partition on $S$ and let $f$ be a clustering function on $S$. A pair of positive numbers $(a, b)$ is $\kappa$-forcing with respect to $f$ if, for every $d \in \mathcal{D}'_S$ that is $(a, b)$-conformant to $\kappa$, we have $f(d) = \kappa$.*

**Lemma 13.23.** *If $f$ is a consistent clustering function on a set $S$, then for any partition $\kappa \in \mathrm{Ran}(f)$ there exist $a, b \in \mathbb{R}_{>0}$ such that the pair $(a, b)$ is $\kappa$-forcing.*

*Proof.* For $\kappa \in \mathrm{Ran}(f)$ there exists $d \in \mathcal{D}'_S$ such that $f(d) = \kappa$. Define the numbers

$$a_{\kappa,d} = \min\{d(x,y) \mid x \neq y, x \equiv_\kappa y\},$$
$$b_{\kappa,d} = \max\{d(x,y) \mid x \not\equiv_\kappa y\}.$$

In other words, $a_{\kappa,d}$ is the smallest $d$ value for two distinct objects that belong to the same $\kappa$-cluster, and $b_{\kappa,d}$ is the largest $d$ value for two objects that belong to different $\kappa$-clusters.

Let $(a,b)$ be a pair of positive numbers such that $a \leq a_{\kappa,d}$ and $b \geq b_{\kappa,d}$. If $d'$ is a definite dissimilarity that is $(a,b)$-conformant to $\kappa$, then $x \equiv_\kappa y$ implies $d'(x,y) \leq a \leq a_{\kappa,d} \leq d(x,y)$ and $x \not\equiv_\kappa y$ implies $d'(x,y) \geq b > b_{\kappa,d} > d(x,y)$, so $d'$ is a $\kappa$-transformation of $d$. By the consistency property of $f$, we have $f(d') = \kappa$. This implies that $(a,b)$ is $\kappa$-forcing. □

**Theorem 13.24.** *If $f$ is a scale-invariant and consistent clustering function on a set $S$, then its range is an antichain in poset $(PART(S), \leq)$.*

*Proof.* This statement is equivalent to saying that, for any scale-invariant and consistent clustering function, no two distinct partitions of $S$ that are values of $f$ are comparable.

Suppose that there are two clusterings, $\kappa_0$ and $\kappa_1$, in the range of a scale-invariant and consistent clustering such that $\kappa_0 < \kappa_1$.

Let $(a_i, b_i)$ be a $\kappa_i$-forcing pair for $i = 0, 1$, where $a_0 < b_0$ and $a_1 < b_1$. Let $a_2$ be a number such that $a_2 \leq a_1$ and choose $\epsilon$ such that

$$0 < \epsilon < \frac{a_0 a_2}{b_0}.$$

By Supplement 27 of Chapter 10 construct a distance $d$ such that

1. for any points $x, y$ that belong to the same block of $\kappa_0$, $d(x,y) \leq \epsilon$;
2. for points that belong to the same cluster of $\kappa_1$ but not to the same cluster of $\kappa_0$, $a_2 \leq d(x,y) \leq a_1$; and
3. for points that do not belong to the same cluster of $\kappa_1$, $d(x,y) \geq b_1$.

The distance $d$ is $(a_1, b_1)$-conformant to $\kappa_1$, and so we have $f(d) = \kappa_1$. Take $\alpha = \frac{b_0}{a_2}$, and define $d' = \alpha d$. Since $f$ is scale-invariant, we have $f(d') = f(d) = \kappa_1$. Note that for points $x, y$ that belong to the same cluster of $\kappa_0$, we have

$$d'(x,y) \leq \frac{\epsilon b_0}{a_2} < a_0,$$

while for points $x, y$ that do not belong to the same cluster of $\kappa_0$ we have

$$d'(x,y) \geq \frac{a_2 b_0}{a_2} \geq b_0.$$

Thus, $d'$ is $(a_0, b_0)$-conformant to $\kappa_0$, and so we must have $f(d') = \kappa_0$. Since $\kappa_0 \neq \kappa_1$, this is a contradiction. □

**Theorem 13.25 (Kleinberg's Impossibility Theorem).** *If $|S| \geq 2$, there is no clustering function that is scale-invariant, rich and consistent.*

*Proof.* If $S$ contains at least two elements, then the poset $(PART(S), \leq)$ is not an antichain. Therefore, this statement is a direct consequence of Theorem 13.24. □

**Theorem 13.26.** *For every antichain $A$ of the poset $(PART(S), \leq)$, there exists a clustering function $f$ that is scale-invariant and consistent such that* $\mathrm{Ran}(f) = A$.

*Proof.* Suppose that $A$ contains more than one partition. We define $f(d)$ as the first partition $\pi \in A$ (in some arbitrary but fixed order) that minimizes the quantity

$$\Phi_d(\pi) = \sum_{x \equiv_\pi y} d(x, y).$$

Note that $\Phi_{\alpha d} = \alpha \Phi_d$. Therefore, $f$ is scale-invariant.

We need to prove that every partition of $A$ is in the range of $f$.

For a partition $\rho \in A$, define $d$ such that $d(x, y) < \frac{1}{|S|^3}$ if $x \equiv_\rho y$ and $d(x, y) \geq 1$ otherwise. Observe that $\Phi_d(\rho) < 1$. Suppose that $\Phi_d(\theta) < 1$. The definition of $d$ means that

$$\Phi_d(\theta) = \sum_{x \equiv_\theta y} d(x, y) < 1,$$

so for all pairs $(x, y) \in \equiv_\theta$ we have $d(x, y) < \frac{1}{|S|^3}$, which means that $x \equiv_\rho y$. Therefore, we have $\pi < \rho$. Since $A$ is an antichain, it follows that $\rho$ must minimize $\Phi_d$ over all partitions of $A$ and, consequently, $f(d) = \rho$.

To verify the consistency of $f$, suppose that $f(d) = \pi$, and let $d'$ be a $\pi$-transformation of $d$. For $\sigma \in PART(S)$, define $\delta(\sigma)$ as $\Phi_d(\sigma) - \Phi_{d'}(\sigma)$. For $\sigma \in A$, we have

$$\begin{aligned}
\delta(\sigma) &= \sum_{x \equiv_\sigma y} (d(x, y) - d'(x, y)) \\
&\leq \sum_{\substack{x \equiv_\sigma y \\ \text{and } x \equiv_\pi y}} (d(x, y) - d'(x, y)) \\
&\qquad \text{(only terms corresponding to pairs in the same cluster are nonnegative)} \\
&\leq \delta(\pi) \\
&\qquad \text{(every term corresponding to a pair in the same cluster is nonnegative).}
\end{aligned}$$

Consequently,

$$\Phi_d(\sigma) - \Phi_{d'}(\sigma) \leq \Phi_d(\pi) - \Phi_{d'}(\pi)$$

or $\Phi_d(\sigma) - \Phi_d(\pi) \leq \Phi_{d'}(\sigma) - \Phi_{d'}(\pi)$. Thus, if $\pi$ minimizes $\Phi_d(\pi)$, then $\Phi_d(\sigma) - \Phi_d(\pi) \geq 0$ for every $\sigma \in A$ and therefore $\Phi_{d'}(\sigma) - \Phi_{d'}(\pi) \geq 0$, which means that $\pi$ also minimizes $\Phi_{d'}(\pi)$. This implies $f(d') = \pi$, which shows that $f$ is consistent. □

## 13.6 Clustering Quality

There are two general approaches for evaluating the quality of a clustering: *unsupervised evaluation*, which measures the cluster cohesion and the separation between clusters, and *supervised evaluation* which measures the extent to which the clustering we analyze matches a partition of the set of objects that is specified by an external labeling of the objects.

### 13.6.1 Object Silhouettes

The *silhouette method* is an unsupervised method for evaluation of clusterings that computes certain coefficients for each object. The set of these coefficients allows an evaluation of the quality of the clustering.

Let $O = \{u_1, \ldots, u_n\}$ be a collection of objects, $d : O \times O \longrightarrow \mathbb{R}_+$ a dissimilarity on $O$, and let $f : O \longrightarrow \{C_1, \ldots, C_k\}$ be a clustering function.

Suppose that $f(u_i) = C_\ell$. The $(f, d)$-*average dissimilarity* is the function $a_{k,d} : O \longrightarrow \mathbb{R}$ given by

$$a_{f,d}(u_i) = \frac{\sum\{d(u_i, u) \mid f(u) = f(u_i) \text{ and } u \neq u_i\}}{|f(u_i)|},$$

that is, the average dissimilarity of $u_i$ to all objects of $f(u_i)$, the cluster to which $u_i$ is assigned.

For a cluster $C$ and an object $u_i$ let

$$d(u_i, C) = \frac{\sum\{d(u_i, u) \mid f(u) = C\}}{|C|},$$

be the average dissimilarity between $u_i$ and the objects of the cluster $C$.

**Definition 13.27.** *Let $f : O \longrightarrow \{C_1, \ldots, C_k\}$ be a clustering function A neighbor of $u_i$ is a cluster $C \neq f(u_i)$ for which $d(u_i, C)$ is minimal.*

In other words, a neighbor of an object $u_i$ is "the second best choice" for a cluster for $u_i$. Let $b : O \longrightarrow \mathbb{R}$ be the function defined by

$$b_{f,d}(u_i) = \min\{d(u_i, C) \mid C \neq f(u_i)\}.$$

If $f$ and $d$ are clear from the context, we shall simply write $a(u_i)$ and $b(u_i)$ instead of $a_{f,d}(u_i)$ and $b_{f,d}(u_i)$, respectively.

**Definition 13.28.** *The silhouette of the object $u_i$ for which $|f(u_i)| \geq 2$ is the number sil$(u_i)$ given by*

$$sil(u_i) = \begin{cases} 1 - \frac{a(u_i)}{b(u_i)} & \text{if } a(u_i) < b(u_i) \\ 0 & \text{if } a(u_i) = b(u_i) \\ \frac{b(u_i)}{a(u_i)} - 1 & \text{if } a(u_i) > b(u_i). \end{cases}$$

*Equivalently, we have*

$$sil(u_i) = \frac{b(u_i) - a(u_i)}{\max\{a(u_i), b(u_i)\}}$$

*for* $u_i \in O$.

*If* $f(u_i) = 1$, *then* $s(u_i) = 0$.

Observe that $-1 \leq sil(u_i) \leq 1$. When $sil(u_i)$ is close to 1, this means that $a(u_i)$ is much smaller than $b(u_i)$ and we may conclude that $u_i$ is well-classified. When $sil(u_i)$ is near 0, it is not clear which is the best cluster for $u_i$. Finally, if $sil(u_i)$ is close to $-1$, the average distance from $u$ to its neighbor(s) is much smaller than the average distance between $u_i$ and other objects that belong to the same cluster $f(u_i)$. In this case, it is clear that $u_i$ is poorly classified.

**Definition 13.29.** *The average silhouette width of a cluster* $C$ *is*

$$sil(C) = \frac{\sum\{sil(u) \mid u \in C\}}{|C|}.$$

*The average silhouette width of a clustering* $\kappa$ *is*

$$sil(\kappa) = \frac{\sum\{sil(u) \mid u \in O\}}{|O|}.$$

The silhouette of a clustering can be used for determining the "optimal" number of clusters. If the average silhouette of the clustering is above 0.7, we have a strong clustering.

### 13.6.2 Supervised Evaluation

Suppose that we intend to evaluate the accuracy of a clustering algorithm $\mathcal{A}$ on a set of objects $S$ relative to a collection of classes on $S$ that forms a partition $\sigma$ of $S$. In other words, we wish to determine the extent to which the clustering produced by $\mathcal{A}$ coincides with the partition determined by the classes.

If the set $S$ is large, the evaluation can be performed by extracting a random sample $T$ from $S$, applying $\mathcal{A}$ to $T$, and then comparing the clustering partition of $T$ computed by $\mathcal{A}$ and the partition of $T$ into the preexisting classes.

Let $\kappa = \{C_1, \ldots, C_m\}$ be the clustering partition of $T$ and let $\sigma = \{K_1, \ldots, K_n\}$ be the partition of $T$ into preexisting classes. The evaluation is helped by $n \times m$-matrix $Q$, where $q_{ij} = |C_i \cap K_j|$ named the *confusion matrix*.

We can use distances associated with the generalized entropy, $d_\beta(\kappa, \sigma)$, to evaluate the distinction between these partitions. This was already observed in [111], who proposed as a measure the cardinality of the symmetric difference

of the sets of pairs of objects that belong to the equivalences that correspond to the two partitions.

Frequently, one uses the conditional entropy

$$\mathcal{H}(\sigma|\kappa) = \sum_{i=1}^{m} \frac{|C_i|}{|T|} \mathcal{H}(\sigma_{C_i}) = \sum_{i=1}^{m} \frac{|C_i|}{|T|} \sum_{j=1}^{n} \frac{|C_i \cap K_j|}{|C_i|} \log_2 \frac{|C_i \cap K_j|}{|C_i|}$$

to evaluate the "purity" of the clusters $C_i$ relative to the classes $K_1, \ldots, K_n$. Low values of this number indicate a high degree of purity.

Some authors [132] define the *purity* of a cluster $C_i$ as $pur_\sigma(C_i) = \max_j \frac{|C_i \cap K_j|}{|C_i|}$ and the purity of the clustering $\kappa$ relative to $\sigma$ as

$$pur_\sigma(\kappa) = \sum_{i=1}^{n} \frac{|C_i|}{|T|} pur_\sigma(C_i).$$

Larger values of the purity indicate better clusterings (from the point of view of the matching with the class partition of the set of objects).

*Example 13.30.* Suppose that a set of 1000 objects consists of three classes of objects $K_1, K_2, K_3$, where $|K_1| = 500$, $|K_2| = 300$, and $|K_1| = 200$. Two clustering algorithms $\mathcal{A}$ and $\mathcal{A}'$ yield the clusterings $\kappa = \{C_1, C_2, C_3\}$ and $\kappa' = \{C'_1, C'_2, C'_3\}$ and the confusion matrices $Q$ and $Q'$, respectively:

| | $K_1$ | $K_2$ | $K_3$ |
|---|---|---|---|
| $C_1$ | 400 | 0 | 25 |
| $C_2$ | 60 | 200 | 75 |
| $C_3$ | 40 | 100 | 100 |

and

| | $K_1$ | $K_2$ | $K_3$ |
|---|---|---|---|
| $C'_1$ | 60 | 0 | 180 |
| $C'_2$ | 400 | 50 | 0 |
| $C'_3$ | 40 | 250 | 20 |

The distances $d_2(\kappa, \sigma)$ and $d_2(\kappa', \sigma)$ are 0.5218 and 0.4204, suggesting that the clustering $\kappa'$ produced by the second algorithm is closer to the partition in classes.

As expected, the purity of the first clustering, 0.7, is smaller than the purity of the second clustering, 0.83.

Another measure of clustering quality, proposed in [112] which applies to objects in $\mathbb{R}^n$ and can be applied, for example, to the clustering that results from the $k$-means method, is the *validity of clustering*. Let $\pi = \{U_1, \ldots, U_k\}$ be a clustering of $N$ objects, $\mathbf{c}_1, \ldots, \mathbf{c}_k$ the centroids of the clusters. The clustering validity is

$$val(\pi) = \frac{sse(\pi)}{N \min_{i<j} d^2(\mathbf{c}_i, \mathbf{c}_j)}.$$

The variety of clustering algorithms is very impressive and it is very helpful to the reader to consult two excellent surveys of clustering algorithms [73, 11] before exploring in depth this domain.

## Exercises and Supplements

1. Apply hierarchical clustering to the data set given in Example 13.4 using the average-link method, the centroid method, and the Ward method. Compare the shapes of the clusters that are formed during the aggregation process. Draw the dendrograms of the clusterings.
2. Using a random number generator, produce $h$ sets of points in $\mathbb{R}^n$ normally distributed around $h$ given points in $\mathbb{R}^n$. Use $k$-means to cluster these points with several values for $k$ and compare the quality of the resulting clusterings.
3. A variant of the $k$-means clustering introduced in [130] is the *bisecting $k$-means algorithm* described below. The parameters are $S$ the set of objects to be clustered, $k$ the desired number of clusters, and $nt$, the number of trial bisections.

   **Algorithm 13.31 bisecting $k$-means**{
   $set_of_clusters = \{S\}$;
   **while** ($|set_of_clusters| < k$)
   extract a cluster $C$ from the $set_of_clusters$;
   $k = 0$;
   for $i = 1$ to $nt$ do
   let $C_{0i}, C_{1i}$ be the two clusters obtained from $C$ by bisecting $C$
   using standard $k$-means ($k = 2$);
   if ($i = 1$) then $s = sse(\{C_{0i}, C_{1i}\})$;
   if ($sse(\{C_{0i}, C_{1i}\}) \leq s$) then
   $k = i$;
   $s = sse(\{C_{0i}, C_{1i}\})$;
   endif;
   endfor;
   add $C_{0k}, C_{1k}$ to $set_of_clusters$;
   **endwhile**
   }

   The cluster $C$ that is bisected may be the largest cluster or the cluster having the largest *sse*.
   Evaluate the time performance of bisecting $k$-means compared with the standard $k$-means and with some variant of a hierarchical clustering.
4. One of the issues that the $k$-means algorithm must confront is that the number of clusters $k$ must be provided as an input parameter. Using clustering validity, design an algorithm that identifies local maxima of validity (as a function of $k$) to provide a basis for a good choice of $k$. See [112] for a solution that applies to image segmentation.
5. Let $B \subseteq \mathbb{R}^n$ be a finite subset of $\mathbb{R}^n$. The *clustering feature* of $B$ is a triple $(p, \mathbf{s}, \mathbf{q})$, where $p = |B|$, $\mathbf{s} = \sum\{\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\}$, and $\mathbf{q} = \big(\sum\{x_1^2 \mid \mathbf{x} \in B\}, \ldots, \sum\{x_n^2 \mid \mathbf{x} \in B\}\big)$. The center of $B$ is $\bar{\mathbf{x}} = \frac{1}{p}\mathbf{s}$, the

*average distance between the center and the members* of $B$ is

$$R_B = \sqrt{\frac{(x - \bar{x})^2}{p}}$$

and the average distance between the members of the clusters is

$$D_B = \sqrt{\frac{\sum\{(\mathbf{u} - \mathbf{v})^2 \mid \mathbf{u}, \mathbf{v} \in B\}}{p(p-1)}}.$$

Prove that $\bar{x}, R_B$, and $D_B$ can be computed starting from the cluster feature.

Let $\mathcal{G} = (V, E)$ be a finite graph. A *graph clustering* [22] is a partition $\kappa = \{C_1, \ldots, C_p\}$ of the set $V$; the clusters are the subgraphs $\mathcal{G}_{C_i} = (C_i, E_{C_i})$ induced by the blocks of $\kappa$. The *intracluster edges* are the edges in $E_\kappa = \bigcup_{i=1}^{p} E_{C_i}$, while the *intercluster edges* are the edges in $E - E_\kappa$. The set of edges between nodes in $C$ and $C'$ is denoted by $E(C, C')$.

6. The quality of a graph clustering $\kappa$ is measured by its *modularity index* $q(\kappa)$ given by

$$q(\kappa) = \sum_{C \in \kappa} \left\{ \frac{|E_\kappa|}{|E|} - \left( \frac{|E(C)| + \sum_{C' \in \kappa} |E(C, C')|}{2|E|} \right)^2 \right\}.$$

Prove that $q(\kappa) = \sum_{C \in \kappa} \left\{ \frac{|E_\kappa|}{|E|} - \left( \frac{\sum_{v \in C} \mathsf{d}(v)}{2|E|} \right)^2 \right\}$ What does it take for a clustering to achieve a high value of the modularity index?

7. Prove that $q(\kappa) \in [-0.5, 1]$ for every clustering of a graph $\mathcal{G}$ and the minimum is achieved when all edges are intercluster edges.
8. Prove that there is always a clustering of a graph $\mathcal{G}$ that has maximum modularity in which each cluster consists of a connected subgraph.
9. Let $\mathcal{G} = (V, E)$ be a bipartite graph with the partition $\pi = \{V_1, V_2\}$ (see Definition 3.7). Prove that $q(\kappa) = -0.5$.
10. Prove that for $k \geq 2$ and for sufficiently large sets of objects, the clustering function $f^g$ introduced in Example 13.18 is not consistent.

    **Solution:** Suppose that $\kappa = \{C_1, C_2, \ldots, C_k\}$ is a partition of $S$ and $d$ is a definite dissimilarity on $S$ such that $d(x, y) = r_i$ if $x \neq y$ and $\{x, y\} \subseteq C_i$ for some $1 \leq i \leq k$ and $d(x, y) = r + a$ if $x$ and $y$ belong to two distinct blocks of $\kappa$, where $r = \max\{r_i | 1 \leq i \leq k\}$ and $a > 0$.

    Suppose that $T$ is a set of $k$ members of $S$. Then, the value of $g(d(x, T))$ is $g(r)$ if the closest member of $T$ is in the same block as $x$ and is $g(r + a)$ otherwise. This means that the smallest value of $\Lambda_d^g(T) = \sum_{x \in C_i} g(d(x, T))$ is obtained when each block $C_i$ contains a member $t_i$ of $T$ for $1 \leq i \leq k$ and the actual value is $\Lambda_d^g(T) = \sum_{i=1}^{k} (|C_i| - 1)r^2 = (|S| - k)r^2$.

Consider now a partition $\kappa' = \{C_1', C_1'', C_2, \ldots, C_k\}$, where $C_1 = C_1' \cup C_1''$ so $\kappa' < \kappa$. Choose $r'$ to be a positive number such that $r' < r$ and define the dissimilarity $d'$ on $S$ such that $d'(x, y) = r'$ if $x \neq y$ and $x \equiv_{\kappa'} y$ and $d'(x, y) = d(x, y)$ otherwise. Clearly, $d'$ is a $\kappa$-transformation of $d$. The minimal value for $\Lambda_d^g(T')$ will be achieved when $T'$ consists of $k + 1$ points, one in each block of $\kappa'$; as a result, the value of the clustering function for $d'$ will be $\kappa' \neq \kappa$, which shows that no clustering function obtained by this technique is consistent.

## Bibliographical Comments

Several general introductions in data mining [132, 130] provide excellent references for clustering algorithms. Basic reference books for clustering algorithms are [72] and [76]. Recent surveys such as [11] and [73] allow the reader to get familiar with current issues in clustering. Cluster features discussed in Exercise 5 were considered in the BIRCH algorithm [150]. Exercises 6–9 contain results obtained in [22].

# Part IV

# Combinatorics

# 14

# Combinatorics

## 14.1 Introduction

Combinatorics is the area of mathematics concerned with counting collections of mathematical objects. Several elementary combinatorial problems were already discussed in Chapter 1, where we counted the number of permutations of a set of size $n$, the number of subsets of size $k$, etc. In this chapter, we present several of the more involved combinatorial techniques that are relevant for data mining.

## 14.2 The Inclusion-Exclusion Principle

Let $A$ and $B$ be two finite sets. It is easy to verify that

$$|A \cup B| = |A| + |B| - |A \cap B|. \tag{14.1}$$

In this section we discuss a generalization of Equality( 14.1) known as the *inclusion-exclusion principle*.

Note that if $U$ and $V$ are two subsets of a finite set $S$ such that $V \subseteq U$, then the function $I$ defined by $I(x) = I_U(x) - I_V(x)$ for $x \in S$ is an indicator function, namely the indicator function of the subset $U - V$ of $S$.

Let $a$ and $b$ be two numbers that belong to the set $\{-1, 1\}$ such that the function $I_{ab}$ defined by

$$I_{ab}(x) = aI_U(x) + bI_V(x)$$

for $x \in S$ is the indicator function of a subset $W$ of the set $S$. Since $I_{ab}(x) \in \{0, 1\}$, the following cases are possible:

1. If $a = b = 1$, then we have $U \cap V = \emptyset$; otherwise (that is, if $x \in U \cap V$) we would have $aI_U(x) + bI_V(x) = 2$ and this would prevent $I_{ab}$ from being an indicator function. Clearly, in this case, $W = U \cup V$.

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_14, © Springer-Verlag London Limited 2008

2. If $a = 1$ and $b = -1$, we must have $I_V(x) \leq I_U(x)$ for every $x \in S$, which implies $V \subseteq U$. Thus, $W = U - V$.
3. The case where $a = -1$ and $b = 1$ is similar to the previous case, and we have $W = V - U$.
4. The case when $a = -1$ and $b = -1$ is possible only if $U = V = \emptyset$. In this case, $W = \emptyset$.

Note that in all these cases we have $|W| = a|U| + b|V|$. This observation is generalized by the following statement.

**Theorem 14.1.** *Let $U_0, \ldots, U_{n-1}$ be $n$ subsets of a finite set $S$, where $n \geq 2$, and let $(a_0, \ldots, a_{n-1}) \in \mathbf{Seq}_n(\{-1, 1\})$ be a sequence of $n$ numbers such that the function $I : S \longrightarrow \{0, 1\}$ defined by*

$$I(x) = a_0 I_{U_0}(x) + \cdots + a_{n-1} I_{U_{n-1}}(x)$$

*for $x \in S$ is the indicator function of a subset $W$ of $S$. Then,*

$$|W| = a_0|U_0| + \cdots + a_{n-1}|U_{n-1}|.$$

*Proof.* If $W$ is a subset of $S$, then $\sum_{x \in S} I_W(x) = |W|$ because for each $x \in S$ its contribution to the sum $\sum_{x \in S} I_W(x)$ is equal to 1 if and only if $x \in W$. Therefore, if $I_W(x) = \sum_{i=0}^{n-1} a_i I_{U_i}(x)$ for $x \in S$, we have

$$\begin{aligned} |W| = \sum_{x \in S} I_W(x) &= \sum_{x \in S} \sum_{i=0}^{n-1} a_i I_{U_i}(x) \\ &= \sum_{i=0}^{n-1} \sum_{x \in S} a_i I_{U_i}(x) \\ &= \sum_{i=0}^{n-1} a_i \sum_{x \in S} I_{U_i}(x) \\ &= \sum_{i=0}^{n-1} a_i |U_i|. \end{aligned}$$

□

**Corollary 14.2 (Principle of Inclusion-Exclusion).** *Let $A_0, \ldots, A_{n-1}$ be $n$ finite sets, where $n \geq 2$. We have*

$$\left| \bigcup_{i=0}^{n-1} A_i \right| = \sum_{0 \leq i \leq n-1} |A_i| - \sum_{0 \leq i_1 < i_2 \leq n-1} |A_{i_1} \cap A_{i_2}| + \sum_{0 \leq i_1 < i_2 < i_3 \leq n-1} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| - \cdots + (-1)^{n+1} |A_0 \cap \cdots \cap A_{n-1}|.$$

*Proof.* Suppose that $A_i \subseteq S$ for $0 \leq i \leq n-1$, where $S$ is a finite set. For $x \in S$, we have $x \notin A = \bigcup_{i=0}^{n-1} A_i$ if and only if $x \notin A_i$ for $0 \leq i \leq n-1$. This is equivalent to writing

$$1 - I_A(x) = (1 - I_{A_{i_0}}(x)) \cdots (1 - I_{A_{i_{n-1}}}(x))$$

for every $x \in S$. This equality is, in turn, equivalent to

$$\begin{aligned}
&I_A(x)\\
&= \sum_{i=0}^{n-1} I_{A_i}(x) - \sum_{0 \leq i_1 < i_2 \leq n-1} I_{A_{i_1}}(x) I_{A_{i_2}}(x)\\
&\quad + \sum_{0 \leq i_1 < i_2 < i_3 \leq n-1} I_{A_{i_1}}(x) I_{A_{i_2}}(x) I_{A_{i_3}}(x) - \cdots + (-1)^{n+1} I_{A_0}(x) \cdots I_{A_{n-1}}(x)\\
&= \sum_{i=0}^{n-1} I_{A_i}(x) - \sum_{0 \leq i_1 < i_2 \leq n-1} I_{A_{i_1} \cap A_{i_2}}(x)\\
&\quad + \sum_{0 \leq i_1 < i_2 < i_3 \leq n-1} I_{A_{i_1} \cap A_{i_2} \cap A_{i_3}}(x) - \cdots + (-1)^{n+1} I_{A_0 \cap \cdots \cap A_{n-1}}(x).
\end{aligned}$$

By applying Theorem 14.1, we obtain the equality of the corollary. □

**Corollary 14.3.** *Let $A_0, \ldots, A_{n-1}$ be $n$ finite sets, where $n \geq 2$, and let $S = \bigcup_{i=0}^{n-1} A_i$. We have*

$$\begin{aligned}
\left| \bigcap_{i=0}^{n-1} A_i \right| &= |S| - \sum_{0 \leq i \leq n-1} |A_i| + \sum_{0 \leq i_1 < i_2 \leq n-1} |A_{i_1} \cap A_{i_2}|\\
&\quad - \sum_{0 \leq i_1 < i_2 < i_3 \leq n-1} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| + \cdots + (-1)^n |A_0 \cap \cdots \cap A_{n-1}|.
\end{aligned}$$

*Proof.* This follows immediately from Corollary 14.2 by observing that

$$\left| \bigcap_{i=0}^{n-1} A_i \right| = |S| - \left| \bigcup_{i=0}^{n-1} A_i \right|.$$

□

It is interesting to observe that the principle of inclusion-exclusion can be obtained also from the Möbius dual inversion theorem. Let $A_0, \ldots, A_{n-1}$ be $n$ finite sets, where $n \geq 2$, $S = \bigcup_{i=0}^{n-1} A_i$, and $I$ be a subset of the set $\{0, \ldots, n-1\}$. The complement of $I$, $\{0, \ldots, n-1\} - I$ is denoted by $\bar{I}$.

Let $B_I$ be the subset of $S$ that consists of those elements that belong to every one of the sets $A_i$ with $i \in I$ and to no other sets. Clearly, we have

$$B_I = \left( \bigcap_{i \in I} A_i \right) \cap \left( \bigcap_{i \in \bar{I}} A_i \right).$$

Note that if $I \neq I'$, then the sets $B_I$ and $B_{I'}$ are disjoint. We claim that

$$\bigcup\{B_J \mid I \subseteq J \subseteq \{0,\ldots,n-1\}\} = \bigcap_{i\in I} A_i. \tag{14.2}$$

If $I \subseteq J$, then $B_J \subseteq \bigcap_{i\in I} A_i$. Therefore,

$$\bigcap_{i\in I} A_i \subseteq \bigcup\{B_J \mid I \subseteq J \subseteq \{0,\ldots,n-1\}\}.$$

Conversely, let $x \in \bigcap_{i\in I} A_i$ and let $J_x = \{j \in \{0,\ldots,n-1\} \mid x \in A_j\}$. It is clear that $I \subseteq J_x$ and that $x \in B_{J_x}$. Therefore, $x \in \bigcup\{B_J \mid I \subseteq J \subseteq \{0,\ldots,n-1\}\}$ and we have the reverse inclusion

$$\bigcap_{i\in I} A_i \subseteq \bigcup\{B_J \mid I \subseteq J \subseteq \{0,\ldots,n-1\}\},$$

which proves Equality (14.2). This allows us to write

$$\left|\bigcap_{i\in I} A_i\right| = \sum\{|B_J| \mid I \subseteq J \subseteq \{0,\ldots,n-1\}\}.$$

Define $f(J)$ as $|B_J|$. The last equality can now be rewritten as

$$\left|\bigcap_{i\in I} A_i\right| = \sum\{f(J) \mid I \subseteq J \subseteq \{0,\ldots,n-1\}\}.$$

By the Möbius dual inversion theorem (Theorem 4.115) applied to the poset $(\mathcal{P}(\{0,\ldots,n-1\}),\subseteq)$, we have

$$f(I) = \sum_{I\subseteq J} (-1)^{|J|-|I|} \left|\bigcap_{i\in J} A_i\right|.$$

For the special case $I = \emptyset$, we have $f(\emptyset) = \left|S - \bigcup_{0\leq i\leq n-1} A_i\right|$ because the intersection of an empty collection of subsets of a set $S$ equals $S$. Thus,

$$\left|S - \bigcup_{0\leq i\leq n-1} A_i\right| = \sum_{J} (-1)^{|J|} \left|\bigcap_{i\in J} A_i\right|,$$

which is equivalent to Corollary 14.3.

*Example 14.4.* Let $n$ be a natural number such that $n \geq 2$. Using the inclusion-exclusion principle we can compute the number $\phi(n)$ of positive integers that are less than $n$ and are relatively prime with $n$; that is, the number of integers $r$ such that $1 \leq r \leq n$ such that $\gcd\{n, r\} = 1$.

Suppose that $n = p_1^{a_1} p_2^{a_2} \cdots p_m^{a_m}$, where $p_1, \ldots, p_m$ are distinct prime numbers and $a_1, \ldots, a_m$ are positive integers. Let $M_i = \{r \in \mathbb{N} \mid r < n \text{ and } p_i | r\}$ for $1 \leq i \leq m$.

It is clear that $|M_i| = \frac{n}{p_i}$ for $1 \leq i \leq m$ and that

$$|M_{i_1} \cap M_{i_2} \cap \cdots \cap M_{i_k}| = \frac{n}{p_{i_1} p_{i_2} \cdots p_{i_k}}$$

for $1 \leq i_1, \ldots, i_k \leq m$.

Note that $r$ is relatively prime with $r$ if and only if $r \notin \bigcup_{i=1}^{m} M_i$.

Thus, the number that we are seeking is $n - |\bigcup_{i=1}^{m} M_i|$. By the inclusion-exclusion principle, we have

$$\begin{aligned}
\phi(n) &= n - \Big| \bigcup_{i=1}^{m} M_i \Big| \\
&= n - \sum_{1 \leq i \leq m} |M_i| + \sum_{1 \leq i_1 < i_2 \leq m} |M_{i_1} \cap M_{i_2}| - \\
&\quad + \cdots + (-1)^m |M_1 \cap \cdots \cap M_m| \\
&= n - \sum_{1 \leq i \leq m} \frac{n}{p_i} + \sum_{1 \leq i_1 < i_2 \leq m} \frac{n}{p_{i_1} p_{i_2}} + \\
&\quad + \cdots + (-1)^m \frac{n}{p_1 p_2 \cdots p_m} \\
&= n \prod_{i=1}^{m} \left(1 - \frac{1}{p_i}\right).
\end{aligned}$$

The function $\phi$ is known as *Euler's function*. It is easy to see that $\phi(2) = 1$, $\phi(3) = 2$, $\phi(4) = 2$, etc. Furthermore, for any prime number $p$, we have $\phi(p) = p - 1$.

## 14.3 Ramsey's Theorem

Data miners should be aware of what is known today as Ramsey theory because this family of combinatorial results establishes that data sets that are sufficiently large contain spurious patterns whose existence is caused by the sheer size of the data set and do not represent "significant" structures from a data mining point of view.

We begin with a set of basic terms of Ramsey theory.

**Definition 14.5.** *Let $C = \{c_1, \ldots, c_k\}$ be a finite set referred to as the* set of colors. *A $C$-coloring of a set $S$ is a mapping $f : S \longrightarrow C$. The set $f^{-1}(c)$ is the set of elements of $S$ colored by $c$.*

*A subset $T$ of $S$ is* monochromatic *in the color $c_i$ if $f(t) = c_i$ for every $t \in T$. A subset $W$ of $S$ is $f$-*monochromatic *if it is monochromatic in some color $c_i$.*

Clearly, every set of the form $f^{-1}(c)$ for a $C$-coloring of $S$ is $f$-monochromatic. Recall that the set of subsets of size $q$ of a set $S$ is denoted by $\mathcal{P}_q(S)$.

**Theorem 14.6.** *Let $S$ be a finite set, $q$ a positive natural number, $f : S \longrightarrow \{c_1, c_2\}$ a coloring of the set $S$ such that every set in $\mathcal{P}_q(S)$ is $f$-monochromatic, and $a_1$ and $a_2$ be two natural numbers not less than* 2.

*There is a number denoted by $\mathcal{R}(a_1, a_2, q)$ such that if $|S| \geq \mathcal{R}(a_1, a_2, q)$, then there is $i \in \{1, 2\}$ and a subset $T$ of $S$ such that $|T| = a_i$ and every subset of $\mathcal{P}_q(T)$ has the color $c_i$.*

*Proof.* We begin by showing that $\mathcal{R}(a_1, q, q) = a_1$ for $a_1 \geq q$. Let $S$ be a set of size $a_1$. One of the following two cases may occur:

`Case 1:` There is a subset $T$ of $S$ of size $q$ that is colored by $c_2$. In this case, the statement holds since the $T$ has only itself as a subset of size $q$.

`Case 2:` There is no subset $T$ of $S$ of size $q$ that is colored by $c_2$. Now all subsets of size $q$ of $S$ have color $c_1$, and if $T$ is a subset of $S$ of size $a_1$, then all its subsets of size $q$ have the color $c_1$ (since they are $q$-subsets of $S$).

This shows that $\mathcal{R}(a_1, q, q) = a_1$ for $a_1 \geq q$; similarly, $\mathcal{R}(q, a_2, q) = a_2$ for $a_2 \geq q$.

The argument is by induction on $q$.

In the basis case, $q = 1$, and we color each element individually. If $S$ is a set of size $a_1 + a_2 - 1$, then we must have either $a_1$ elements colored $c_1$ or $a_2$ elements colored $c_2$ since otherwise, the set $S$ would have no more than $a_1 + a_2 - 2$ elements.

For the inductive step, suppose the theorem holds for $q - 1$. Now we act by induction on $p = a_1 + a_2$. The basis case, where $a_1 = a_2 = q$, is included in the previous discussion.

Suppose that the theorem holds for $a_1 + a_2 - 1$, and let $b_1 = \mathcal{R}(a_1 - 1, a_2, q)$ and $b_2 = \mathcal{R}(a_1, a_2 - 1, q)$. Let $S$ be a set whose size is at least $\mathcal{R}(b_1, b_2, q-1)+1$, and suppose that all its $q$-subsets are colored by $c_1$ or $c_2$. If $s$ is a fixed element of $S$, then any set $U \in \mathcal{P}_q(S)$ such that $s \in U$ yields a subset $U - \{s\}$ of size $q-1$ of the set $S'$, where $S' = S - \{s\}$ is colored in the same color as $U$. Thus, we obtain a coloring of the $q-1$ subsets of the set $S - \{s\}$ that contains at least $\mathcal{R}(b_1, b_2, q - 1)$ elements. By the inductive hypothesis, there is either a subset $V$ of $S'$ such that $|V| = b_1 = \mathcal{R}(a_1 - 1, a_2, q)$ and all its $q - 1$ subsets have color $c_1$ or there is an subset $W$ of $S'$ such that $|W| = b_2 = \mathcal{R}(a_1, a_2 - 1, q)$ and all its $q - 1$ subsets have color $c_2$.

The first case yields a coloring of $S$ in which the $q$-subsets of $S$ obtained by adding $s$ to the $(q - 1)$-subsets of $S'$ are colored in $c_1$. By the definition of $\mathcal{R}(a_1 - 1, a_2, q)$, there exists either a subset $T_1$ of $S'$ that has $a_1 - 1$ elements whose $q$-subsets are colored $c_1$ or a $a_2$-subset $T_2$ of $S$ whose $q$-subsets are colored $c_2$. The statement follows in the first situation by observing that $T_1 \cup \{s\}$ has $a_1$ elements. The second situation requires no further argument.

The second case is treated similarly. □

**Corollary 14.7.** *We have the inequality*

$$\mathcal{R}(a_1, a_2, q) \leq \mathcal{R}(\mathcal{R}(a_1 - 1, a_2, q), \mathcal{R}(a_1, a_2 - 1, q)) + 1$$

*for every* $q \geq 1$ *and* $a_1, a_2$ *such that* $a_1, a_2 \geq q$.

*Proof.* The inequality follows immediately from the proof of Theorem 14.6. □

In the proof of Ramsey's theorem, we shall use the preliminary result contained in Theorem 14.6.

**Theorem 14.8 (Ramsey's Theorem).** *Let* $S$ *be a finite set,* $q$ *a positive natural number,* $f : S \longrightarrow \{c_1, \ldots, c_k\}$ *a coloring of the set* $S$ *such that every set in* $\mathcal{P}_q(S)$ *is* $f$*-monochromatic, and* $\boldsymbol{a} = (a_1, \ldots, a_k)$ *a sequence of* $k$ *positive natural numbers such that* $a_i \geq q$ *for* $1 \leq i \leq k$.

*There is a number denoted by* $\mathsf{Ramsey}(\boldsymbol{a}, q)$ *such that if* $|S| \geq \mathsf{Ramsey}(\boldsymbol{a}, q)$, *then there exists a number* $i$, $1 \leq i \leq k$, *and a set* $T$ *with* $|T| = a_i$ *such that every subset of* $\mathcal{P}_q(T)$ *has the color* $c_i$.

*Proof.* This time the proof is by induction on $k$, the number of colors. The basis case, $k = 2$, was discussed in Theorem 14.6. We have $\mathsf{Ramsey}((a_1, a_2), q) = \mathcal{R}(a_1, a_2, q)$.

Suppose the statement holds for $k - 1$ colors.

Let $S$ be a set such that $|S| \geq \mathsf{Ramsey}((\mathsf{Ramsey}((a_1, \ldots, a_{k-1}), q), a_k), q)$ and let $f : S \longrightarrow \{c_1, \ldots, c_k\}$ be a coloring of $S$ using $k$ colors. Define the coloring $g : S \longrightarrow \{c_0, c_k\}$ by

$$g(x) = \begin{cases} c_0 & \text{if } f(x) \in \{c_1, \ldots, c_{k-1}\}, \\ c_k & \text{if } f(x) = c_k, \end{cases}$$

for $x \in S$. Using the coloring $g$, every $q$-subset of $S$ that was colored $c_1, \ldots, c_{k-1}$ will receive the color $c_0$ and every $q$-subset of $S$ colored $c_k$ will remain colored by $c_k$. By the two-color case of Theorem 14.6, either there is a subset $T$ such that $|T| = \mathsf{Ramsey}((a_1, \ldots, a_{k-1}), q)$ whose $q$-subsets are colored $c_0$ or a subset $U$ such that $|U| = a_k$ whose $q$-subsets are colored $c_k$. Since $f$ colors the $q$-subsets of $T$ in any of the colors $c_1, \ldots, c_{k-1}$, the theorem follows immediately from the inductive hypothesis. □

**Corollary 14.9.** *We have the inequality*

$$\mathsf{Ramsey}((a_1, \ldots, a_k), q) \leq \mathsf{Ramsey}((\mathsf{Ramsey}((a_1, \ldots, a_{k-1}), q), a_k), q)$$

*for every* $q \geq 1$ *and* $a_i$ *such that* $a_i \geq q$ *for* $1 \leq i \leq k$.

*Proof.* This result follows from the proof of Ramsey's theorem. □

Note that $\mathsf{Ramsey}((\underbrace{2,\ldots,2}_{k}),1) = k+1$. Indeed, if we color the elements of a set $S$ using $|S|+1$ colors, then there is a subset $T$ of $S$ that contains two elements colored with the same color. This is a well known combinatorial fact known as the *pigeonhole principle*.

A beautiful application of Theorem 14.8 is known as the Erdös-Szekeres theorem. We need the following preliminary observation.

**Lemma 14.10.** *Let $P$ be a set of points in $\mathbb{R}^2$. If every four-point subset of $P$ is a convex polygon, then the set $P$ itself is a convex polygon.*

*Proof.* This is a direct consequence of Theorem B.12. □

**Theorem 14.11 (The Erdös-Szekeres Theorem).** *For every number $n \in \mathbb{N}$, $n \geq 3$, there exists a number $E(n)$ such that any set $P$ of points in the plane such that $|P| = E(n)$ and no three points of $P$ are collinear contains an $n$-point convex polygon.*

*Proof.* A four-point subset of $P$ may or may not be a convex polygon. Thus, the four-point subsets may be colored with two colors: $c_1$ for convex polygons and $c_2$ for the other four-point sets.

Another key observation is that every five-point set in $\mathbb{R}^2$ such that no three points are collinear contains a four-point convex polygon.

Choose $E(n) = \mathsf{Ramsey}((n,5),4)$, which involves coloring all sets in $\mathcal{P}_4(P)$ with the colors $c_1$ and $c_2$. Note that by Klein's theorem, (Theorem B.14), no five point set can be colored in $c_2$ (which would mean that none of its four-point sets is convex). Therefore, there exists an $n$-element set $K$ that can be colored by $c_1$, and, by Lemma 14.10, the set $K$ is convex. □

Ramsey's theorem can be used to derive interesting properties of graphs. Indeed, if $G = (V, E)$ is a graph, then $E \subseteq \mathcal{P}_2(V)$, and Ramsey functions of the form $\mathsf{Ramsey}(\mathbf{a}, 2)$ yield lower bounds of the cardinality of the vertex sets that guarantee the existence of subgraphs having monochromatic sets of edges and containing a number $a_i$ of vertices for some $i$, $1 \leq i \leq n$, and $\mathbf{a} = (a_1, \ldots, a_n)$.

## 14.4 Combinatorics of Partitions

Let $S$ be a set having $n$ elements. We are interested in the number of partitions of $S$ that have $m$ blocks. We begin by counting the number of onto functions of the form $f : A \longrightarrow B$, where $|A| = n$, $|B| = m$, and $n \geq m$.

**Lemma 14.12.** *Let $A$ and $B$ be two sets, where $|A| = n$, $|B| = m$, and $n \geq m$. The number of surjective functions from $A$ to $B$ is given by*

$$\sum_{j=0}^{m-1}(-1)^j\binom{m}{j}(m-j)^n.$$

*Proof.* There are $m^n$ functions of the form $f : A \longrightarrow B$.

We begin by determining the number of functions that are not surjective. Suppose that $B = \{b_1, \ldots, b_m\}$, and let $F_j = \{f : A \longrightarrow B \mid b_j \notin f(A)\}$ for $1 \leq j \leq m$. A function is not surjective if it belongs to one of the sets $F_j$. Thus, we need to evaluate $|\bigcup_{j=1}^m F_j|$. By using the inclusion-exclusion principle, we can write

$$\left|\bigcup_{j=1}^m F_j\right| = \sum_{j_1=1}^m |A_{j_1}| - \sum_{j_1,j_2=1}^m |A_{j_1} \cap A_{j_2}| + \sum_{j_1,j_2,j_3=1}^m |A_{j_1} \cap A_{j_2} \cap A_{j_3}| - \cdots - +(-1)^m|A_1 \cap A_2 \cap \cdots \cap A_m|.$$

Note that the set $|A_{j-1} \cap A_{j_2} \cap \cdots \cap A_{j_p}|$ is actually the set of functions defined on $A$ with values in the set $B - \{y_{j_1}, y_{j_2}, \ldots, y_{j_p}\}$, and there are $(m-p)^n$ such functions. Since there are $\binom{m}{p}$ choices for the set $\{j_1, j_2, \ldots, j_p\}$, it follows that there are

$$\binom{m}{1}(m-1)^n - \binom{m}{2}(m-2)^n + \binom{m}{1}(m-3)^n - \cdots + (-1)^m\binom{m}{m-1}$$

functions that are not surjective.

Thus, we can conclude that there are

$$\sum_{j=0}^{m-1}(-1)^j\binom{m}{j}(m-j)^n = m^n - \binom{m}{1}(m-1)^n + \binom{m}{2}(m-2)^n - \cdots + (-1)^{m-1}\binom{m}{m-1}$$

surjective functions from $A$ to $B$. □

**Theorem 14.13.** *The number of partitions of a set $S$ that have $m$ blocks $(m \leq n)$ is given by*

$$\frac{1}{m!}\sum_{j=0}^{m-1}(-1)^j\binom{m}{j}(m-j)^n.$$

*Proof.* Note that there are $m!$ distinct onto functions that have the same kernel partition. Indeed, given a surjective function $f : A \longrightarrow B$, one can obtain a function $g$ that has the same partition as $f$ by defining $g(a) = p(f(a))$, where $p$ is a permutation of the set $B$, that is, a bijection $p : B \longrightarrow B$.

Since there are $m!$ such bijections, it follows that the number of partitions is $\frac{1}{m!}\sum_{j=0}^{m-1}(-1)^j\binom{m}{j}(m-j)^n$. □

The numbers $S(n,m)$ defined by

$$S(n,m)=\frac{1}{m!}\sum_{j=0}^{m-1}(-1)^j\binom{m}{j}(m-j)^n$$

for $m,n\in\mathbb{N}$ and $m\le n$ are known as the *Stirling numbers of the second kind*.

So far, we have examined partitions of sets. Next we consider partitions of natural numbers.

**Definition 14.14.** *An* integral partition of $n$ *is a nonincreasing sequence* $\boldsymbol{k}=(k_1,\ldots,k_\ell)$ *of positive integers such that* $\sum_{i=1}^{\ell}k_i=n$.

*The set of integral partitions of* $n$ *is denoted by* $IP_n$*; the set of integral partitions of* $n$ *that consist of* $\ell$ *components is denoted by* $IP_n(\ell)$.

*Example 14.15.* The sequence $\mathbf{k}=(5,5,3,2,2,2,1,1)$ is an integral partition of 21.

We can regard an integral partition of $n$ as a multiset $P$ on the set $\{1,2,\ldots,n\}$, where $P(k)$ is the number of entries in the sequence $\mathbf{k}$ of Definition 14.14 that equal $k$.

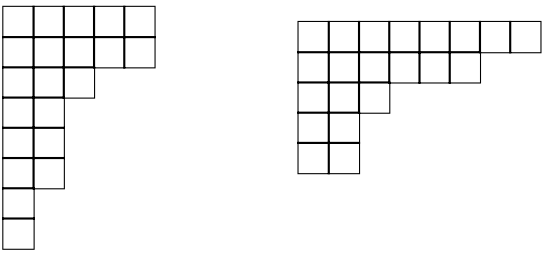
*Example 14.16.* The integral partition $(5,5,3,2,2,2,1,1)\in IP_{21}$ defines the multiset $P$ on the set $\{1,\ldots,21\}$ given by

$$P(k)=\begin{cases}2 & \text{if } k=1 \text{ or } k=5,\\ 3 & \text{if } k=2,\\ 1 & \text{if } k=3,\\ 0 & \text{in every other case.}\end{cases}$$

An integral partition $\mathbf{k}$ can be represented graphically by a *Ferrers diagram* that consists of a sequence of rows of squares such that each component $k$ of $\mathbf{k}$ corresponds to a row of $k$ cells in the diagram.

*Example 14.17.* The Ferrers diagram of $(5,5,3,2,2,2,1,1)$ of integer 21 is shown in Figure 14.1(a).

Starting from the Ferrers diagram of $\mathbf{k}\in IP_n$, we can derive a new integral partition $\mathbf{k}'\in IP_n$ by exchanging the rows of the diagram with its columns. The new integral partition $\mathbf{k}'$ is called the *conjugate integral partition* of $\mathbf{k}$. The Ferrers diagram of the conjugate partition $\mathbf{k}'$ of $\mathbf{k}$ (where $\mathbf{k}$ is the integral partition defined in Example 14.17) is shown in Figure 14.1(b).



(a) (b)

**Fig. 14.1.** Ferrers diagrams.

**Theorem 14.18.** *The number of integral partitions in $IP_n$ where the largest component is $\ell$ equals $IP_n(\ell)$, the number of integral partitions on $n$ with $\ell$ components.*

*Proof.* This statement follows immediately by observing that the function $f : IP_n \longrightarrow IP_n$ that maps $\mathbf{k}$ into its conjugate $\mathbf{k}'$ is a bijection and the image under $f$ of an integral partition that has $\ell$ components is an integral partition whose largest component is $\ell$. □

## 14.5 Combinatorics of Collections of Sets

**Definition 14.19.** *A* Sperner system *is a collection of sets $\mathcal{C}$ such that $X, Y \in \mathcal{C}$ and $X \neq Y$ implies $X \not\subseteq Y$.*

*If $\mathcal{C}$ is a Sperner system and $\mathcal{C} \subseteq \mathcal{P}(S)$, then we say that $\mathcal{C}$ is a* Sperner system on the set $S$.

The next theorem presents an inequality known as the LYM inequality, an acronym of the names of the mathematicians whose work is related to it (Lubell, Yamamoto, and Meshalkin [91, 145, 98]).

**Theorem 14.20 (The LYM Inequality).** *Let $\mathcal{C}$ be a Sperner system on a finite set $S$ such that $|S| = n$. Define the function $c : \{0, 1, \ldots, n\} \longrightarrow \mathbb{N}$ by $c(k) = |\{X \in \mathcal{C} \mid |X| = k\}|$ for $0 \leq k \leq n$. We have*

$$\sum_{k=0}^{n} \frac{c(k)}{\binom{n}{k}} \leq 1.$$

*Proof.* Let $(x_1, \ldots, x_n)$ be one of the $n!$ permutations of the set $S$. For $U \in \mathcal{C}$ define the set $P_U$ as

$$P_U = \{(x_1, \ldots, x_n) \mid \{x_1, \ldots, x_m\} = U, \text{ where } m = |U|\}.$$

Since $\mathcal{C}$ is a Sperner system, we have $P_U \cap P_V = \emptyset$ for $U \neq V$ and $U, V \in \mathcal{C}$. The number of permutations in $P_U$ is $|U|!(n - |U|)!$, so

$$\sum_{U \in \mathcal{C}} |U|!(n - |U|)! \leq n!.$$

Using the definition of the function $c$, we have

$$\sum_{k=0}^{n} c(k)|k|!(n - |k|)! \leq n!,$$

which immediately yields the desired inequality. □

The next statement is known as *Sperner's theorem* and was obtained in [127] using a different approach (outlined in Supplement 21).

**Corollary 14.21.** *Let $S$ be a finite set such that $|S| = n$. If $\mathcal{C}$ is a Sperner system on $S$, then*

$$|\mathcal{C}| \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

*Proof.* The largest value of the binomial coefficient $\binom{n}{k}$ is achieved when $k = \lfloor \frac{n}{2} \rfloor$. Therefore, we have

$$\sum_{k=0}^{n} c(k)|k|!(n - |k|)! \geq \frac{\sum_{i=0}^{n} c(k)}{\binom{n}{\lfloor \frac{n}{2} \rfloor}} = \frac{|\mathcal{C}|}{\lfloor \frac{n}{2} \rfloor}.$$

By the LYM inequality we obtain $|\mathcal{C}| \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}$. □

The Ahlswede-Daykin inequality involves functions defined on sets and collections of sets. We use the operations "$\vee$" and "$\wedge$" between collections of sets introduced in Definition 1.20.

Let $\mathcal{E}$ be a collection of subsets of a set $S$ and let $\phi : \mathcal{P}(S) \longrightarrow \mathbb{R}$ be a function. We define a new function (denoted by the same letter $\phi$) on the set of all collections of subsets of $S$ (that is, on $\mathcal{P}(\mathcal{P}(S))$) as

$$\phi(\mathcal{E}) = \sum \{\phi(E) \mid E \in \mathcal{E}\}.$$

This definition allows us to formulate a powerful combinatorial inequality.

**Theorem 14.22 (The Ahlswede-Daykin Inequality).** *Let $S$ be a set such that $S \neq \emptyset$ and let*

$$\alpha, \beta, \gamma, \delta : \mathcal{P}(S) \longrightarrow \mathbb{N}$$

*be four functions that satisfy the inequality*

$$\alpha(A)\beta(B) \leq \gamma(A \cup B)\delta(A \cap B)$$

*for $A, B \in \mathcal{P}(S)$. For all collections $\mathcal{A}, \mathcal{B}$ of subsets of $S$, we have*

$$\alpha(\mathcal{A})\beta(\mathcal{B}) \leq \gamma(\mathcal{A} \vee \mathcal{B})\delta(\mathcal{A} \wedge \mathcal{B}).$$

*Proof.* The argument is by induction on $n = |S|$, where $n \geq 1$. For the base case, $|S| = 1$, we have $\mathcal{P}(S) = \{\emptyset, S\}$. Thus, we can write

$$\alpha(\emptyset)\beta(\emptyset) \leq \gamma(\emptyset)\delta(\emptyset), \tag{14.3}$$
$$\alpha(\emptyset)\beta(S) \leq \gamma(S)\delta(\emptyset), \tag{14.4}$$
$$\alpha(S)\beta(\emptyset) \leq \gamma(S)\delta(\emptyset), \tag{14.5}$$
$$\alpha(S)\beta(S) \leq \gamma(S)\delta(S). \tag{14.6}$$

Since $\mathcal{C}, \mathcal{B} \subseteq \{\emptyset, S\}$, we need to analyze the following cases.

| Case | $\mathcal{A}$ | $\mathcal{B}$ | $\mathcal{A} \vee \mathcal{B}$ | $\mathcal{A} \wedge \mathcal{B}$ |
|---|---|---|---|---|
| I | $\{\emptyset\}$ | $\{\emptyset\}$ | $\{\emptyset\}$ | $\{\emptyset\}$ |
| II | $\{\emptyset\}$ | $\{S\}$ | $\{S\}$ | $\{\emptyset\}$ |
| III | $\{S\}$ | $\{\emptyset\}$ | $\{S\}$ | $\{\emptyset\}$ |
| IV | $\{S\}$ | $\{S\}$ | $\{S\}$ | $\{S\}$ |
| V | $\{\emptyset\}$ | $\{\emptyset, S\}$ | $\{\emptyset, S\}$ | $\{\emptyset\}$ |
| VI | $\{S\}$ | $\{\emptyset, S\}$ | $\{S\}$ | $\{\emptyset, S\}$ |
| VII | $\{\emptyset, S\}$ | $\{\emptyset\}$ | $\{\emptyset, S\}$ | $\{\emptyset\}$ |
| VIII | $\{\emptyset, S\}$ | $\{S\}$ | $\{S\}$ | $\{\emptyset, S\}$ |
| IX | $\{\emptyset, S\}$ | $\{\emptyset, S\}$ | $\{\emptyset, S\}$ | $\{\emptyset, S\}$ |

We discuss only case IX; the remaining cases are similar and are left to the reader. The inequality that we need to prove,

$$(\alpha(\emptyset) + \alpha(S))(\beta(\emptyset) + \beta(S)) \leq (\gamma(\emptyset) + \gamma(S))(\delta(\emptyset) + \delta(S))$$

follows immediately by adding Inequalities (14.3) to (14.6).

Suppose that the inequality holds for sets containing $m$ elements, and let $S = \{s_0, \ldots, s_{m-1}, s_m\}$ be a set of size $m+1$. Define $U = \{s_0, \ldots, s_{m-1}\}$ and $V = \{s_m\}$.

The mappings $\alpha_1, \beta_1, \gamma_1, \delta_1 : \mathcal{P}(U) \longrightarrow \mathbb{N}$ are defined by

$$\alpha_1(C) = \sum\{\alpha(A) \mid A \in \mathcal{A} \text{ and } A \cap U = C\},$$
$$\beta_1(C) = \sum\{\alpha(B) \mid B \in \mathcal{B} \text{ and } B \cap U = C\},$$
$$\gamma_1(C) = \sum\{\gamma(E) \mid E \in \mathcal{A} \vee \beta \text{ and } E \cap U = C\},$$
$$\delta_1(C) = \sum\{\delta(F) \mid F \in \mathcal{A} \wedge \mathcal{B} \text{ and } F \cap U = C\}.$$

Observe that

$$\begin{aligned}\alpha_1(\mathcal{P}(U)) &= \sum\{\alpha_1(C) \mid C \in \mathcal{P}(U)\}\\ &= \sum_{C\in\mathcal{P}(U)} \sum\{\alpha(A) \mid A \in \mathcal{A} \text{ and } A \cap U = C\}\\ &= \sum\{\alpha(A) \mid A \in \mathcal{A}\}\\ &= \alpha(\mathcal{A}).\end{aligned}$$

Similarly, $\beta_1(\mathcal{P}(U)) = \beta(\mathcal{B})$, $\gamma_1(\mathcal{P}(U)) = \gamma(\mathcal{A} \vee \mathcal{B})$, and $\delta_1(\mathcal{P}(U)) = \gamma(\mathcal{A} \wedge \mathcal{B})$.

Let $R \in \mathcal{P}(V)$. We have either $R = \emptyset$ or $R = \{s_m\}$.

Let $C, D \in \mathcal{P}(U)$ and let $E = C \cup D$ and $F = C \cap D$. Define the mappings $\alpha_2^C, \beta_2^D, \gamma_2^E, \delta_2^F : \mathcal{P}(V) \longrightarrow \mathbb{N}$ by

$$\alpha_2^C(R) = \begin{cases}\alpha(R \cup C) & \text{if } R \cup C \in \mathcal{A},\\ 0 & \text{otherwise,}\end{cases}$$

$$\beta_2^D(R) = \begin{cases}\beta(R \cup D) & \text{if } R \cup D \in \mathcal{B},\\ 0 & \text{otherwise,}\end{cases}$$

$$\gamma_2^E(R) = \begin{cases}\gamma(R \cup E) & \text{if } R \cup E \in \mathcal{A} \vee \mathcal{B},\\ 0 & \text{otherwise,}\end{cases}$$

$$\delta_2^F(R) = \begin{cases}\delta(R \cup F) & \text{if } R \cup F \in \mathcal{A} \wedge \mathcal{B},\\ 0 & \text{otherwise.}\end{cases}$$

We have $\alpha_1(C) = \alpha_2^C(\mathcal{P}(V))$ for every $C \subseteq U$. Indeed,

$$\begin{aligned}\alpha_2^C(\mathcal{P}(V)) &= \alpha_2^C(\emptyset) + \alpha_2^C(\{s_m\})\\ &= \begin{cases}\alpha(C) + \alpha(C \cup \{s_m\}) & \text{if } C \in \mathcal{A} \text{ and } C \cup \{s_m\} \in \mathcal{A},\\ \alpha(C) & \text{if } C \in \mathcal{A} \text{ and } C \cup \{s_m\} \notin \mathcal{A},\\ \alpha(C \cup \{s_m\}) & \text{if } C \notin \mathcal{A} \text{ and } C \cup \{s_m\} \in \mathcal{A},\\ 0 & \text{otherwise.}\end{cases}\\ &= \alpha_1(C).\end{aligned}$$

Similar arguments show that $\beta_1(D) = \beta_2^D(\mathcal{P}(V))$ for $D \in \mathcal{P}(U)$, $\gamma_1(E) = \gamma_2^E(\mathcal{P}(V))$ for $E \in \mathcal{P}(U)$, and $\delta_1(F) = \delta_2^F(\mathcal{P}(V))$ for $F \in \mathcal{P}(U)$.

We claim that $\alpha_2^C(R)\beta_2^D(Q) \leq \gamma_2^E(R \cup Q)\delta_2^F(R \cap Q)$ for all $R, Q \in \mathcal{P}(V)$.

If $\alpha_2^C(R)\beta_2^D(Q) = 0$ the inequality obviously holds.

Now suppose that $\alpha_2^C(R)\beta_2^D(Q) \neq 0$, that is, $R \cup C \in \mathcal{A}$ and $Q \cup D \in \mathcal{B}$ and $\alpha_2^C(R)\beta_2^D(Q) = \alpha(R \cup C)\beta(Q \cup D)$. Note that

$$(R \cup C) \cup (Q \cup D) = (R \cup Q) \cup (C \cup D) = (R \cup Q) \cup E \in \mathcal{A} \vee \mathcal{B}$$

and

$$\begin{aligned}(R \cup C) \cap (Q \cup D) &= (R \cap Q) \cup (R \cap D) \cup (C \cap Q) \cup (C \cap D) \\ &= (R \cap Q) \cup (C \cap D) = (R \cap Q) \cup F \in \mathcal{A} \wedge \mathcal{B}\end{aligned}$$

because $R \cap D = C \cap Q = \emptyset$ (since $R, Q \in \mathcal{P}(V)$ and $C, D \in \mathcal{P}(U)$). Thus,

$$\gamma_2^E(R \cup Q)\delta_2^F(R \cap Q) = \gamma(R \cup Q \cup E)\delta((R \cap Q) \cup F).$$

By the defining property of $\alpha, \beta, \gamma$, and $\delta$, we have $\alpha(R \cup C)\beta(Q \cup D) \leq \gamma(R \cup Q \cup E)\delta((R \cap Q) \cup F)$, which yields the inequality

$$\alpha_2^C(R)\beta_2^D(Q) \leq \gamma_2^E(R \cup Q)\delta_2^F(R \cap Q)$$

for all $R, Q \in \mathcal{P}(V)$.

The inductive hypothesis (for $n = 1$) implies

$$\begin{aligned}&\alpha_1(C)\beta_1(D) \\ &= \alpha_2^C(\mathcal{P}(V))\beta_2^D(\mathcal{P}(V)) \leq \gamma_2^E(\mathcal{P}(V))\delta_2^F(\mathcal{P}(V)) \\ &= \gamma_1(C \cup D)\delta_1(C \cap D).\end{aligned}$$

Again applying the inductive hypothesis, we can write

$$\alpha(\mathcal{A})\beta(\mathcal{B}) = \alpha_1(\mathcal{P}(U))\beta_1(\mathcal{P}(U)) \leq \gamma_1(\mathcal{P}(U))\delta_1(\mathcal{P}(U)) = \gamma(\mathcal{A} \vee \mathcal{B})\delta(\mathcal{A} \wedge \mathcal{B}).$$

□

**Corollary 14.23.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two collections of subsets of $S$. In this case,*

$$|\mathcal{A}| \cdot |\mathcal{B}| \leq |\mathcal{A} \vee \mathcal{B}| \cdot |\mathcal{A} \wedge \mathcal{B}|.$$

*Proof.* In the Ahlswede-Daykin inequality, choose $\alpha, \beta, \gamma, \delta : \mathcal{P}(S) \longrightarrow \mathbb{N}$ such that $\alpha(C) = \beta(C) = \gamma(C) = \delta(C) = 1$ for $C \in \mathcal{P}(S)$. The required inequality follows immediately. □

**Definition 14.24.** *A* hereditary collection of sets *is a collection $\mathcal{I}$ such that $C \in \mathcal{I}$ and $D \subseteq C$ implies $D \in \mathcal{I}$.*

*A* dually hereditary collection of sets *is a collection of sets $\mathcal{F}$ such that $C \in \mathcal{F}$ and $C \subseteq D$ implies $D \in \mathcal{F}$.*

Note that if $\mathcal{I}$ is a hereditary family of subsets of a set $S$, then $\mathcal{P}(S) - \mathcal{I}$ is a dually hereditary family of subsets; similarly, if $\mathcal{F}$ is a dually hereditary family of subsets of $S$, then $\mathcal{P}(S) - \mathcal{F}$ is a hereditary family.

**Theorem 14.25.** *Let $\mathcal{I}$ and $\mathcal{I}'$ be two hereditary families of sets. Then,*

$$\mathcal{I} \vee \mathcal{I}' = \mathcal{I} \cap \mathcal{I}'.$$

*Proof.* Let $C \in \mathcal{I} \vee \mathcal{I}'$. Then, $C = A \cap B$, where $A \in \mathcal{I}$ and $B \in \mathcal{I}'$. Since $C \subseteq A$ and $C \subseteq B$, the hereditary character of $\mathcal{I}$ and $\mathcal{I}'$ implies that $C \in \mathcal{I}$ and $C \in \mathcal{I}'$, so $C \in \mathcal{I} \cap \mathcal{I}'$. □

**Theorem 14.26.** *Let $\mathcal{F}$ and $\mathcal{F}'$ be two dual hereditary families of sets. Then,*

$$\mathcal{F} \wedge \mathcal{F}' = \mathcal{F} \cap \mathcal{F}'.$$

*Proof.* Let $C \in \mathcal{F} \wedge \mathcal{F}'$. Then, $C = A \cup B$, where $A \in \mathcal{F}$ and $B \in \mathcal{F}'$. Since $A \subseteq C$ and $B \subseteq C$, the dual hereditary character of $\mathcal{F}$ and $\mathcal{F}'$ implies that $C \in \mathcal{F}$ and $C \in \mathcal{F}'$, so $C \in \mathcal{F} \cap \mathcal{F}'$. □

The inequality contained by the next corollary is known as *Kleitman's inequality.*

**Corollary 14.27.** *If $\mathcal{I}$ is a hereditary family and $\mathcal{F}$ is a dual hereditary family of subsets of a finite set $S$, then $|\mathcal{I}| \cdot |\mathcal{F}| \geq 2^{|S|} \cdot |\mathcal{I} \cap \mathcal{F}|$.*

*Proof.* Note that $\mathcal{I}' = \mathcal{P}(S) - \mathcal{F}$ is a hereditary family. By Corollary 14.23, we have

$$\begin{aligned} |\mathcal{I}| \cdot |\mathcal{I}'| &\leq |\mathcal{I} \vee \mathcal{I}'| \cdot |\mathcal{I} \wedge \mathcal{I}'| \\ &= |\mathcal{I} \cap \mathcal{I}'| \cdot |\mathcal{I} \wedge \mathcal{I}'| \\ &\quad \text{(by Theorem 14.25)} \end{aligned}$$

Note that $|\mathcal{I}'| = 2^{|S|} - |\mathcal{F}|$. Thus, we can write

$$\begin{aligned} |\mathcal{I}| \cdot \left(2^{|S|} - |\mathcal{F}|\right) &= |\mathcal{I} - (\mathcal{I} \cap \mathcal{F})| \cdot |\mathcal{I} \wedge \mathcal{I}'| \\ &= (|\mathcal{I}| - |\mathcal{I} \cap \mathcal{F}|) \cdot |\mathcal{I} \wedge \mathcal{I}'| \\ &\leq 2^{|S|} \cdot (|\mathcal{I}| - |\mathcal{I} \cap \mathcal{F}|), \end{aligned}$$

which gives the desired inequality. □

If $\phi$ is logarithmic supramodular on $S$ (see Definition 8.53), then, by the Ahlswede-Daykin inequality, we have

$$\phi(\mathcal{A})\phi(\mathcal{B}) \leq \phi(\mathcal{A} \vee \mathcal{B})\phi(\mathcal{A} \wedge \mathcal{B}),$$

for every $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}(S)$ and $\phi(\mathcal{E}) = \sum_{E \in \mathcal{E}} \phi(E)$, for $\mathcal{E} \subseteq \mathcal{P}(S)$.

## Exercises and Supplements

1. A *derangement* is a permutation $f : \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$ such that $f(i) \neq i$ for $1 \leq i \leq n$. Denote by $E_{\{i_1 \cdots i_k\}}$ the set of permutations $f$ of $PERM_n$ such that $f(i_p) = i_p$ for $1 \leq p \leq k$.
   a) Prove that $|E_{\{i_1 \cdots i_k\}}| = (n-k)!$.
   b) By applying the inclusion-exclusion principle, prove that the number of derangements is

$$D_n = n! \left(1 - \frac{1}{1!} + \frac{1}{2!} + \cdots + (-1)^n \frac{1}{n!}\right).$$

c) Use a combinatorial argument to prove that

$$n! = \sum_{j=0}^{n} \binom{n}{j} D_{n-j}.$$

2. Let $S$ be a set, $\mathcal{C}$ be a collection of subsets of $S$ and $T$ be a subset of $S$. Denote by $d(\mathcal{C}, T)$ the collection of sets in $\mathcal{C}$ that are disjoint from $T$, $\{C \in \mathcal{C} \mid C \cap T = \emptyset\}$. If $c_k$ is the number of ways to choose a subcollection $\mathcal{D}$ of $\mathcal{C}$ such that $|\mathcal{D}| = k$ and $\bigcup \mathcal{S} = S$, then prove that $c_k = \sum_{T \subseteq S} (-1)^{|T|} |d(\mathcal{C}, T)|^k$.

   **Solution:** Note that there are $|d(\mathcal{C}, T)|^k$ ways to pick $k$ sets $C_1, \ldots, C_k$ of $\mathcal{C}$ that are disjoint from $T$. If $\bigcup C_i = S$, then at least one of these sets must intersect $T$, which implies $T = \emptyset$. Thus, every cover contributes only to the term $(-1)^0 |d(\mathcal{C}, \emptyset)|^k$.

   If $\bigcup C_i = V \subset S$, then the $C_i$ contribute to every term corresponding to $T = S - V$, so the total contribution of $C_1, \ldots, C_k$ is $\sum_{T \subseteq S-V} (-1)^T$, and this sum equals 0 because every nonempty set has an equal number of even and odd-sized subsets. Thus, only the collection of sets that cover the entire set $S$ contributes to $c_k$.
3. Let $P$ be a subset of $\{1, \ldots, 2n\}$ such that $|P| = n + 1$. Prove that there exists a pair of numbers in $P \times P$ whose components are relatively prime numbers.
4. Prove that $\mathcal{R}(m, p, q) \leq \binom{m + p - 2}{m - 1}$ for $m \geq 2$ and $p \geq 2$.
5. Let $\mathcal{G} = (\mathbb{N}, E)$ be a complete graph having $\mathbb{N}$ as its set of vertices and $E = (m, n) \in \mathcal{P}_2(\mathbb{N}) \mid m \neq n$.
   a) If $f : E \longrightarrow \{c_1, c_2\}$ is a two-color coloring of the edges of $\mathcal{G}$, prove that there exists an infinite complete subgraph of $\mathcal{G}$ that is monochromatic.
   b) Extend this statement to an $r$-color coloring, where $r \geq 3$.

   **Solution:** Define the sequence of infinite subsets $T_0, T_1, \ldots$ of $\mathbb{N}$ as follows. The initial set is $T_0 = \mathbb{N}$. Suppose $T_i$ is defined. Choose $n_i \in T_i$, and let $U_{ij} = \{n \in T - \{n_i\} \mid f(n_i, n) = c_j\}$ for $j = 1, 2$. At least one of $U_{i1}, U_{i2}$ is infinite, and $T_{i+1}$ is chosen as one of $U_{i1}, U_{i2}$ that is infinite.

   If $i \leq \min\{j, k\}$, then $n_j \in T_j \subset T_{i+1}$ and $n_k \in T_k \subset T_{i+1}$, which implies $f(n_i, n_j) = f(n_i, n_k)$ because of the definition of the sets $T_i$. Let $U = \{n_0, n_1, \ldots\}$ and let $g : U \longrightarrow \{c_1, c_2\}$ be given by $g(n_i) = f(n_i, n_j)$ for $i < j$. It is clear that $g$ is well-defined. Since $U$ is an infinite set, at least one of the subsets $g^{-1}(c_1), g^{-1}(c_2)$ is infinite. Let $W$ be one of these subsets that is infinite. Then, for $n_l, n_k \in W$, where $l < k$, we have $g(n_l) = g(n_k) = c$ and therefore $f(n_l, n_k) = c$. Thus, the subgraph induced by $U$ is monochromatic.
6. Let $\mathbf{n} = (n_0, n_1, \ldots) \in \mathbf{Seq}_\infty(\mathbb{N})$ be a sequence of natural numbers. Prove that $\mathbf{n}$ contains a subsequence that is either strictly increasing, strictly decreasing, or constant. Extend this result to countable, totally ordered sets.

**Hint:** Consider the complete graph on the set $\{n_0, n_1, \ldots\}$, and color each edge $(n_i, n_j)$ with $i < j$ with red if $n_i < n_j$, blue if $n_i > n_j$, and white if $n_i = n_j$.

7. Let $\mathbf{a} = (a_1, \ldots, a_n)$ be a sequence in $\mathbf{Seq}_n(\mathbb{N})$ such that $r \leq \min a_i$ and let $p = \max a_i$. If $\mathbf{p} = (p, \ldots, p) \in \mathbf{Seq}_n(\mathbb{N})$, prove that $\mathsf{Ramsey}(\mathbf{p}, r) \geq \mathsf{Ramsey}(\mathbf{a}, r)$.
8. The *left shift* and the *right shift* on $PERM_n$ are the mappings $lshift, rshift : PERM_n \longrightarrow PERM_n$ defined by

$$lshift(a_1, a_2, \ldots, a_n) = (a_2, \ldots, a_n, a_1),$$
$$rshift(a_1, a_2, \ldots, a_n) = (a_n, a_1 \ldots, a_{n-1}),$$

for every $(a_1, \ldots, a_n) \in PERM_n$, respectively.
   a) Prove that *lshift* and *rshift* are inverse to each other.
   b) Two permutations $f, g \in PERM_n$ are equivalent, $f \equiv g$, if there exists an integer $k$ such that $lshift^{(k)}(f) = g$. Here $h^{(k)}$ denotes the $k^{\text{th}}$ iteration of $h$.
   Prove that "$\equiv$" is an equivalence on $PERM_n$.
9. Prove that $S(n, 2) = 2^{n-1} - 1$ for $n \geq 2$.
10. Prove that

$$S(n, m) = mS(n-1, m) + S(n-1, m-1),$$

for $1 \leq m \leq n$, using a combinatorial argument.
11. The number of partitions of a set having $n$ elements is denoted by $B_n$ and is known as the $n^{\text{th}}$ *Bell number*. Clearly, $B_n = \sum_{m=0}^{n} S(n, m)$. Prove that:

$$B_n = \sum_{m=0}^{n-1} \binom{n-1}{m} B_m.$$

12. Prove that if $0 \leq p \leq (n+1)! - 1$, then there exists a unique sequence $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbf{Seq}(nn)$ such that $0 \leq a_i \leq i$ and $p = a_1 \cdot 1! + a_2 \cdot 2! + \cdots + a_n \cdot n!$.
13. Consider the polynomial $[x]_n = x(x-1)\cdots(x-n+1)$. The coefficients of this polynomial

$$[x]_n = s(n, n)x^n + s(n, n-1)x^{n-1} + \cdots + s(n, i)x^i + \cdots + s(n, 0)$$

are known as *the Stirling numbers of the first kind*.
Prove that

$$s(n, 0) = 0,$$
$$s(n, n) = 1,$$
$$s(n+1, k) = s(n, k-1) - ns(n, k).$$

**Hint:**: To prove the last equality, note that $[x]_{n+1} = [x]_n(x - n)$ and seek the coefficient of $x_k$ in both sides.

14. The Stirling numbers of the second kind occur in an equality related to the one used to define the Stirling numbers of the first kind.
Prove that every polynomial $x^n$ can be written as

$$x^n = \sum_{i=1}^{n} S(n,i)[x]_i.$$

**Solution:** Let $A$ and $B$ be two finite sets such that $|A| = n$ and $|B| = m$. There are $m^n$ functions $f : A \longrightarrow B$. These functions can be classified depending on the size of their range $f(A)$. If $g : A \longrightarrow B$ is a function such that $|g(A)| = j$, then $g$ can be regarded as a surjection from $A$ to $g(A)$. Since there are $j!S(n,j)$ such onto functions and there are $\binom{m}{j}$ subsets of $B$ that have $j$ elements, we can write

$$\begin{aligned} m^n &= \sum_{j=1}^{m} \binom{m}{j} j!S(n,j) \\ &= m(m-1)\cdots(m-j+1)S(n,j). \end{aligned}$$

15. Prove the inequality

$$S(n,i-1)S(n,i+1) \leq (S(n,i))^2$$

for $1 \leq i \leq n-1$.

16. Consider the equation $x_1 + \cdots + x_p = n$, where $n \geq 1$. Prove that:
a) the number of solutions in natural numbers is $\binom{n+p-1}{p-1}$;
b) the number of solutions in positive integers is $\binom{n-1}{p-1}$.
**Solution:** Let $S = \{a, b\}$ be a two-element set and let $\mathbf{a} = (a, \ldots, a) \in \mathbf{Seq}_n(S)$ be a sequence of length $n$. Let $\mathbf{a}' \in \mathbf{Seq}_{n+p-1}$ be the sequence obtained from $\mathbf{a}$ by inserting $p-1$ elements $b$. The number of $a$ symbols between any two consecutive $b$s yields a solution in natural numbers if adjacent $b$ symbols are allowed and there are $\binom{n+p-1}{p-1}$ configurations of $\mathbf{a}'$, each corresponding to a choice of $p-1$ positions out of a total of $n+p-1$ possible places. Further, any solution of the equation can be obtained in this manner. The second part follows immediately from the first part.

17. Prove that $|IP_n(\ell)|$ equals the number of solutions of the equation $y_1 + \cdots + y_\ell = n - \ell$ such that $y_1 \geq y_2 \geq \cdots \geq y_\ell \geq 0$. Infer that $|IP_n(\ell)| = \sum_{k=1}^{\ell} |IP_{n-\ell}(k)|$.

18. Prove that the number of Ferrers diagrams that can be placed in an $m \times n$ rectangle is $\binom{m+n}{n}$.

19. A partition $\mathbf{k} \in IP_n$ is *self-conjugate* if $\mathbf{k}' = \mathbf{k}$. Using Ferrers diagrams prove that the number of self-conjugate partitions in $IP_n$ equals the number of partitions of $n$ into distinct odd numbers.

20. Let $S$ be a set having $n$ elements.

a) A collection $\mathcal{A}$ of subsets of $S$ has the *intersecting property* if $A, B \in \mathcal{A}$ implies $A \cap B \neq \emptyset$. Prove that if $\mathcal{A}$ has the intersecting property, then $|\mathcal{A}| \leq 2^{n-1}$.

b) Prove that there are collections of subsets of $S$ that have the intersecting property and contain $2^{n-1}$ subsets.

**Solution:** For Part 1, note that, for any subset $B$, at most one of the sets $B$, $S - B$ may belong to $\mathcal{A}$. Therefore, $\mathcal{A}$ may not contain more than half of the members of $\mathcal{P}(S)$, so $|\mathcal{A}| \leq 2^{n-1}$.

A collection of sets that answers Part (b) is the set of subsets of $S$ that contain a fixed element $a$ of $S$. Clearly, there are $2^{n-1}$ such sets.

21. Prove Sperner's theorem using Supplement 54 of Chapter 1.

**Solution:** Let $\mathcal{C}$ be a Sperner system on a finite set $S$ such that $|S| = n$ and let $c : \{0, 1, \ldots, n\} \longrightarrow \mathbb{N}$ be defined by $c(k) = |\{X \in \mathcal{C} \mid |X| = k\}$ for $0 \leq k \leq n$. Suppose that $i_0 = \min\{i \mid 1 \leq i \leq n \mid c(i) > 0\} < \frac{n-1}{2}$. By Part (e) of Supplement 54 in Chapter 1, for each of the sets $X$ in $\mathcal{C}$ with $|X| = i_0$, there exists a set $X'$ in the shade of $\{X \in \mathcal{C} \mid |X| = i_0\}$ such that $|X'| = i_0 + 1$ and $X' \notin \mathcal{C}$. By replacing each $X$ with the corresponding $X'$, we obtain a new Sperner system with the same number of sets as $\mathcal{C}$. The process is repeated until a Sperner system $\mathcal{C}'$ that contains no sets with fewer than $\frac{n-1}{2}$ elements is obtained and $\mathcal{C}' = |\mathcal{C}|$. Thus each set in $\mathcal{C}'$ has at least $\frac{n+1}{2}$ elements. Now the process is reversed using the $\Delta\mathcal{C}'$ by replacing every set of $\mathcal{C}'$ of size $\frac{n+1}{2}$ by a set of size $\lfloor \frac{n}{2} \rfloor$. The Sperner system $\mathcal{C}''$ has the same size as $\mathcal{C}$ and consists of sets of size $\frac{n+1}{2}$, so $|\mathcal{C}| \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}$.

22. Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of subsets of a set $S$. In Definition 1.20, we introduced the collection $\mathcal{D} - \mathcal{C}$ as $\mathcal{C} - \mathcal{D} = \{U - V \mid U \in \mathcal{C}, V \in \mathcal{D}\}$. Prove that $|\mathcal{C} - \mathcal{C}| \geq |\mathcal{C}|$.

**Solution:** Let $\mathcal{D}$ be a collection of subsets of a set $S$ and let $\mathcal{D}' = \{S - D \mid D \in \mathcal{D}\}$. Observe that $|\mathcal{D}| = |\mathcal{D}'|$.

If $\mathcal{C}$ and $\mathcal{D}$ are two collections of subsets of $S$, we have $|\mathcal{C} \vee \mathcal{D}'| = |(\mathcal{C} \vee \mathcal{D}')'| = |\mathcal{C}' \wedge \mathcal{D}|$. By the previous observation and by Corollary 14.23, we can write

$$\begin{aligned} |\mathcal{C}| \cdot |\mathcal{D}| &= |\mathcal{C}| \cdot |\mathcal{D}'| \leq |\mathcal{C} \vee \mathcal{D}'| \cdot |\mathcal{C} \wedge \mathcal{D}'| \\ &= |\mathcal{C}' \wedge \mathcal{D}| \cdot |\mathcal{C} \wedge \mathcal{D}'| = |\mathcal{D} - \mathcal{C}| \cdot |\mathcal{C} - \mathcal{D}|. \end{aligned}$$

If we now choose $\mathcal{D} = \mathcal{C}$, the previous inequality yields $|\mathcal{C}|^2 \leq |\mathcal{C} - \mathcal{C}|^2$, which gives the desired inequality.

23. Prove that if $\mathcal{C}$ and $\mathcal{D}$ are hereditary or dually hereditary families of subsets of a finite set $S$, then $|\mathcal{C}| \cdot |\mathcal{D}| \leq 2^n \cdot |\mathcal{C} \cap \mathcal{D}|$.

24. Let $I$ be a set of items and $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set. Recall that in Section 7.6 we introduced the function $tids_T : \mathcal{P}(I) \longrightarrow \mathcal{P}(\{1, \ldots, n)$ by $tids_T(H) = \{k \in \{1, \ldots, n\} \mid H \subseteq T(k)\}$ for any item set $H$.

a) Prove that if $L, J \subseteq I$, $J \subseteq L$, and $L - J = \{i_1, \ldots, i_p\}$, then $tids_T(L) = \bigcap_{\ell=1}^{p} tids_T(J \cup \{i_\ell\})$.
b) Let $F_J^L$ be the number $F_J^L = |\{(h, J') \mid J' \subseteq T(h) \text{ and } J' \cap L = J|$. Prove that

$$F_J^L = \left| \bigcup_{k=1}^{p} tids_T(J \cup \{i_k\}) \right| - |tids_T(J)|.$$

c) By applying the Inclusion-Exclusion Principle, prove that

$$suppcount(L) - (-1)^p F_J^L = \sum_{J \subseteq J' \subset L} (-1)^{|L-J'|+1} suppcount(J').$$

## Bibliographical Comments

Supplement 2 was obtained in [15]. The inequality from Supplement 22 was obtained in [96]. Exercise 24 contains a result of T. Calders [26].

There are several well-known and comprehensive references on combinatorics that contain rich collections of ideas [128, 129] and [57].

# 15

# The Vapnik-Chervonenkis Dimension

## 15.1 Introduction

The concept of the Vapnik-Chervonenkis dimension of a collection of sets was introduced in [139] and independently in [117]. Its main interest for data mining is related to one of the basic models of machine learning, the probabilistic approximately correct learning paradigm as was shown in [16]. The subject is of great interest to probability theorists interested in empirical processes [41, 108].

## 15.2 The Vapnik-Chervonenkis Dimension

**Definition 15.1.** *Let $U$ be a set, $K$ be a subset of $U$, and $\mathcal{C}$ be a collection of subsets of $U$, $\mathcal{C} \subseteq \mathcal{P}(U)$. If the trace of $\mathcal{C}$ on $K$, $\mathcal{C}_K$ equals $\mathcal{P}(K)$, then we say that $K$ is* shattered by $\mathcal{C}$.

*The* Vapnik-Chervonenkis dimension *of the collection $\mathcal{C}$ (called the VC-dimension for brevity) is the largest cardinality of a set $K$ that is shattered by $\mathcal{C}$ and is denoted by VCD($\mathcal{C}$).*

If $U$ is a finite set, then the trace of a collection $\mathcal{C} = \{C_1, \ldots, C_p\}$ of subsets of $U$ on a subset $K$ of $U$ can be presented in an intuitive, tabular form. Suppose, for example, that $U = \{u_1, \ldots, u_n\}$, and let $\theta = (T_{\mathcal{C}}, u_1 u_2 \cdots u_n, \mathbf{r})$ be a table, where $\mathbf{r} = (t_1, \ldots, t_p)$. The domain of each of the attributes $u_i$ is the set $\{0, 1\}$.

Each tuple $t_k$ corresponds to a set $C_k$ of $\mathcal{C}$ and is defined by

$$t_k[u_i] = \begin{cases} 1 & \text{if } u_i \in C_k, \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq i \leq n$. Then, $\mathcal{C}$ shatters $K$ if the content of the projection $\mathbf{r}[K]$ consists of $2^{|K|}$ distinct rows.

D.A. Simovici, C. Djeraba, *Mathematical Tools for Data Mining,*
DOI: 10.1007/978-1-84800-201-2_15, © Springer-Verlag London Limited 2008

*Example 15.2.* Let $U = \{u_1, u_2, u_3, u_4\}$ and let $\mathcal{C}$ be the collection of subsets of $U$ given by

$$\mathcal{C} = \{\{u_2, u_3\}, \{u_1, u_3, u_4\}, \{u_2, u_4\}, \{u_1, u_2\}, \{u_2, u_3, u_4\}\}.$$

The tabular representation of $\mathcal{C}$ is

$T_{\mathcal{C}}$

| $u_1$ | $u_2$ | $u_3$ | $u_4$ |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 | 0 | 1 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |

The set $K = \{u_1, u_3\}$ is shattered by the collection $\mathcal{C}$ because

$$\mathbf{r}[K] = ((0,1), (1,1), (0,0), (1,0), (0,1))$$

contains the all four necessary tuples $(0,1)$, $(1,1)$, $(0,0)$, and $(1,0)$. On the other hand, it is clear that no subset $K$ of $U$ that contains at least three elements can be shattered by $\mathcal{C}$ because this would require $\mathbf{r}[K]$ to contain at least eight tuples. Thus, $VCD(\mathcal{C}) = 2$.

**Theorem 15.3.** *Let $U$ be a finite nonempty set and let $\mathcal{C}$ be a collection of subsets of $U$. If $d = VCD(\mathcal{C})$, then $2^d \leq |\mathcal{C}| \leq (|U| + 1)^d$.*

*Proof.* If $\mathcal{C}$ shatters a finite set $K$ of size $d$, then $\mathcal{C}_K$ must contain at least $2^d$ sets. Therefore, $2^d \leq |\mathcal{C}|$.

The argument for the inequality $|\mathcal{C}| \leq (|U| + 1)^d$ is by induction on $|U|$.

The basis step, $|U| = 1$, is immediate.

Suppose that the statement holds for sets of size at most $n$ and let $U$ be a set such that $|U| = n + 1$. Select an element $x_0 \in U$ and define the set $U_1 = U - \{x_0\}$ and the collection of sets

$$\mathcal{C}_0 = \{D \in \mathcal{C} \mid D = E \cup \{x_0\} \text{ for some } E \in \mathcal{C} \text{ and } x_0 \notin E\}.$$

Let $\mathcal{C}_1 = \mathcal{C} - \mathcal{C}_0$.

The sets of $\mathcal{C}_0$ are distinct on $U_1$; in other words, if $D \cap U_1 = D' \cap U_1$ for $D, D' \in \mathcal{C}_1$, then $D = D'$, as can be seen immediately.

Let $\mathcal{C}_0' = \{D - \{x_0\} \mid D \in \mathcal{C}_1\}$. The definition of $\mathcal{C}_0$ implies that $\mathcal{C}_0' \subseteq \mathcal{C}$ and that $\mathcal{C}_0' \subseteq \mathcal{P}(U_1)$. It is clear that the collections $\mathcal{C}_0'$ and $\mathcal{C}_0$ have the same cardinality.

If $\mathcal{C}'_0$ shatters a subset $S$ of $U_1$ with $|S| = d$, then $S \cup \{x_0\}$ will be shattered by $\mathcal{C}_0$, and therefore by $\mathcal{C}$. Since this is not possible (because $|S \cup \{x_0\}| = d+1$ and $VCD(\mathcal{C}) = d$), it follows that $VCD(\mathcal{C}'_0) \leq d-1$. Thus, by the inductive hypothesis, $|\mathcal{C}'_0| \leq (|U_1|+1)^{d-1}$.

If $D \in \mathcal{C}_1$, then $D \in \mathcal{C}$ and we have two cases:

1. $x_0 \notin D$ and therefore $D \subseteq U_1$, or
2. $x_0 \in D$ and $D - \{x_0\} \notin \mathcal{C}$.

Consider the collection $\mathcal{C}'_1 = \{D - \{x_0\} \mid D \in \mathcal{C}_1\}$. The mapping $\ell : \mathcal{C}_1 \longrightarrow \mathcal{C}'_1$ defined by $\ell(D) = D - \{x_0\}$ is a bijection. It is immediate that $\ell$ is a surjection; thus, we need to show only that $\ell$ is injective.

Suppose that $D - \{x_0\} = D' - \{x_0\}$ for $D, D' \in \mathcal{C}_1$. If $x_0 \notin D$, then $D = D' - \{x_0\}$. Suppose that $x_0 \in D'$. In this case, $D' - \{x_0\} \notin \mathcal{C}$ and this contradicts the fact that $D \in \mathcal{C}$. Thus, $x_0 \notin D'$ and so, $D = D'$.

If $x_0 \in D$, then $D - \{x_0\} \notin \mathcal{C}$. Thus, $D' - \{x_0\} \notin \mathcal{C}$, which happens only if $x_0 \in D'$. Thus, $D = D'$, so $\ell$ is indeed a bijection. This allows us to conclude that $\mathcal{C}'_1$ is a collection of concepts on $U_1$ that has the same cardinality as $\mathcal{C}_1$.

If $\mathcal{C}'_1$ shatters a subset $Z$ of $U_1$ of size larger than $d$, then $\mathcal{C}_1$ shatters $Z$, which is not possible. Thus, $VCD(\mathcal{C}'_1) \leq d$. By the inductive hypothesis, $|\mathcal{C}'_1| \leq (|U_1|+1)^d$, which means that $|\mathcal{C}_1| \leq (|U_1|+1)^d$.

We conclude that

$$\begin{aligned} |\mathcal{C}| &= |\mathcal{C}_0| + |\mathcal{C}_1| \leq (|U_1|+1)^{d-1} + (|U_1|+1)^d \\ &\leq (|U_1|+1)^{d-1}(|U_1|+2) \leq (|U_1|+2)^d = (|U|+1)^d. \end{aligned}$$

□

It is clear that every collection of sets shatters the empty set. Also, if $\mathcal{C}$ shatters a set of size $n$, then it shatters a set of size $p$, where $p \leq n$.

For a collection of sets $\mathcal{C}$ and for $m \in \mathbb{N}$, let $\mathcal{C}[m]$ be the number

$$\mathcal{C}[m] = \max\{|\mathcal{C}_K| \mid |K| = m\}.$$

This is the largest number of distinct subsets of a set having $m$ elements that can be obtained as intersections of the set with members of $\mathcal{C}$. In general, $\mathcal{C}[m] \leq 2^m$; however, if $\mathcal{C}$ shatters a set of size $m$, then $\mathcal{C}[m] = 2^m$.

**Definition 15.4.** *A* Vapnik-Chervonenkis class *(or a* VC class*) is a collection* $\mathcal{C}$ *of sets such that* $VCD(\mathcal{C})$ *is finite.*

*Example 15.5.* Let $\mathbb{R}$ be the set of real numbers and let $\mathcal{S}$ be the collection of sets $\{(-\infty, t) \mid t \in \mathbb{R}\}$. We claim that any singleton is shattered by $\mathcal{S}$. Indeed, if $S = \{x\}$ is a singleton, then $\mathcal{P}(\{x\}) = \{\emptyset, \{x\}\}$. Thus, if $t \geq x$, we have $(-\infty, t) \cap S = \{x\}$; also, if $t < x$, we have $(-\infty, t) \cap S = \emptyset$, so $\mathcal{S}_S = \mathcal{P}(S)$.

There is no set $S$ with $|S| = 2$ that can be shattered by $\mathcal{S}$. Indeed, suppose that $S = \{x, y\}$, where $x < y$. Then, any member of $\mathcal{S}$ that contains $y$ includes the entire set $S$, so $\mathcal{S}_S = \{\emptyset, \{x\}, \{x, y\}\} \neq \mathcal{P}(S)$. This shows that $\mathcal{S}$ is a VC class and $VCD(\mathcal{S}) = 1$.

*Example 15.6.* Consider the collection $\mathcal{I} = \{[a, b] \mid a, b \in \mathbb{R}, a \leq b\}$ of closed intervals. We claim that $VCD(\mathcal{I}) = 2$. To justify this claim, we need to show that there exists a set $S = \{x, y\}$ such that $\mathcal{I}_S = \mathcal{P}(S)$ and no three-element set can be shattered by $\mathcal{I}$.

For the first part of the statement, consider the intersections

$$\begin{aligned}
&[u, v] \cap S = \emptyset, \text{ where } v < x,\\
&[x - \epsilon, \tfrac{x+y}{2}] \cap S = \{x\},\\
&[\tfrac{x+y}{2}, y] \cap S = \{y\},\\
&[x - \epsilon, y + \epsilon] \cap S = \{x, y\},
\end{aligned}$$

which show that $\mathcal{I}_S = \mathcal{P}(S)$.

For the second part of the statement, let $T = \{x, y, z\}$ be a set that contains three elements. Note that any interval that contains $x$ and $z$ also contains $y$, so it is impossible to obtain the set $\{x, z\}$ as an intersection between an interval in $\mathcal{I}$ and the set $T$.

*Example 15.7.* Let $\mathcal{H}$ be the collection of closed half-planes in $\mathbb{R}^2$, that is, the collection of sets of the form

$$\{x = (x_1, x_2) \in \mathbb{R}^2 \mid ax_1 + bx_2 - c \geq 0, a \neq 0 \text{ or } b \neq 0\}.$$

We claim that $VCD(\mathcal{H}) = 3$.

Let $P, Q, R$ be three points in $\mathbb{R}^2$ such that they are not located on the same line. Each line in Figure 15.1 is marked with the sets it defines; thus, it is clear that the family of hyperplanes shatters the set $\{P, Q, R\}$, so $VCD(\mathcal{H})$ is at least 3.

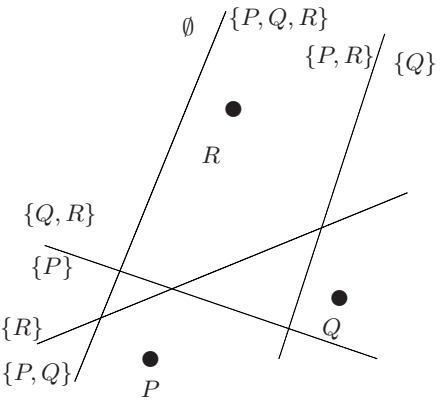



**Fig. 15.1.** Three-point sets can be shattered by half-planes.

To complete the justification of the claim we need to show that no set that contains at least four points can be shattered by $\mathcal{H}$.

Let $\{P, Q, R, S\}$ be a set that contains four points such that no three points of this set are collinear. If $S$ is located inside the triangle $P, Q, R$, then every half-plane that contains $P, Q, R$ will contain $S$, so it is impossible to separate the subset $\{P, Q, R\}$. Thus, we may assume that no point is inside the triangle formed by the remaining three points (see Figure 15.2). Note that any half-plane that contains two diagonally opposite points, for example, $P$ and $R$, will contain either $Q$ or $S$, which shows that it is impossible to separate the set $\{P, R\}$. Thus, no set that contains four points may be shattered by $\mathcal{H}$, so $VCD(\mathcal{H}) = 3$.

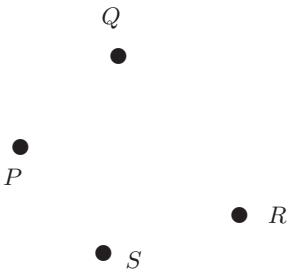



**Fig. 15.2.** A four-point set cannot be shattered by half-planes.

*Example 15.8.* Let $\mathbb{R}^2$ be equipped with a system of coordinates and let $\mathcal{R}$ be the set of rectangles whose sides are parallel with the axes $x$ and $y$. Each such rectangle has the form $[x_0, x_1] \times [y_0, y_1]$.

There is a set $S$ with $|S| = 4$ that is shattered by $\mathcal{R}$. Indeed, let $S$ be a set of four points in $\mathbb{R}^2$ that contains a unique "northernmost point" $P_n$, a unique "southernmost point" $P_s$, a unique "easternmost point" $P_e$, and a unique "westernmost point" $P_w$. If $L \subseteq S$ and $L \neq \emptyset$, let $R_L$ be the smallest rectangle that contains $L$. For example, we show the rectangle $R_L$ for the set $\{P_n, P_s, P_e\}$ in Figure 15.3.

On the other hand, this collection cannot shatter a set of points that contains at least five points. Indeed, let $S$ be a set of points such that $|S| \geq 5$ and, as before, let $P_n$ be the northernmost point, etc. If the set contains more than one "northernmost" point, then we select exactly one to be $P_n$. Then, the rectangle that contains the set $K = \{P_n, P_e, P_s, P_w\}$ contains the entire set $S$, which shows the impossibility of separating the set $K$.

If a collection of sets $\mathcal{C}$ is not a VC class (that is, if the Vapnik-Chervonenkis dimension of $\mathcal{C}$ is infinite), then $\mathcal{C}[m] = 2^m$ for all $m \in \mathbb{N}$. However, we shall prove that if $VCD(\mathcal{C}) = d$, then $\mathcal{C}[m]$ is bounded asymptotically by a polynomial of degree $d$.

For $n, k \in \mathbb{N}$ and $0 \leq k \leq n$ define the number $\binom{n}{\leq k}$ as

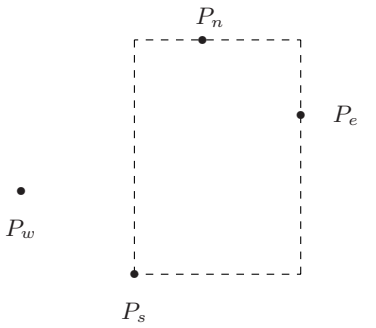



**Fig. 15.3.** Rectangle that separates the set $\{P_n, P_s, P_e\}$.

$$\binom{n}{\leq k} = \sum_{i=0}^{k} \binom{n}{k}$$

**Theorem 15.9.** *Let $\phi : \mathbb{N}^2 \longrightarrow \mathbb{N}$ be the function defined by*

$$\phi(d,m) = \begin{cases} 1 & \text{if } m = 0 \text{ or } d = 0 \\ \phi(d, m-1) + \phi(d-1, m-1) & \text{otherwise.} \end{cases}$$

*We have*

$$\phi(d,m) = \binom{m}{\leq d}$$

*for $d, m \in \mathbb{N}$.*

*Proof.* The argument is by strong induction on $s = i + m$. The base case, $s = 0$, implies $m = 0$ and $d = 0$, and the equality is immediate. Suppose that the equality holds for $\phi(d', m')$, where $d' + m' < d + m$. We have

$$\begin{aligned} \phi(d,m) &= \phi(d, m-1) + \phi(d-1, m-1) \\ &\quad \text{(by definition)} \\ &= \textstyle\sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\ &\quad \text{(by inductive hypothesis)} \\ &= \textstyle\sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} \\ &\quad \text{(since } \tbinom{m-1}{-1} = 0\text{)} \\ &= \textstyle\sum_{i=0}^{d} \left( \binom{m-1}{i} + \binom{m-1}{i-1} \right) \\ &= \textstyle\sum_{i=0}^{d} \binom{m}{i} = \binom{m}{\leq d}, \end{aligned}$$

which gives the desired conclusion. □

**Theorem 15.10 (Sauer-Shelah Theorem).** *If $\mathcal{C}$ is a collection of subsets of $S$ that is a VC-class such that $VCD(\mathcal{C}) = d$, then $\mathcal{C}[m] \leq \phi(d,m)$ for $m \in \mathbb{N}$, where $\phi$ is the function defined in Theorem 15.9.*

*Proof.* The argument is by strong induction on $s = d + m$. For the base case, $s = 0$ we have $d = m = 0$ and this means that the collection $\mathcal{C}$ shatters only the empty set. Thus, $\mathcal{C}[0] = |\mathcal{C}_\emptyset| = 1$, and this implies $\mathcal{C}[0] = 1 = \phi(0, 0)$.

Suppose that the statement holds for pairs $(d', m')$ such that $d' + m' < s$ and let $\mathcal{C}$ be a collection of subsets of $S$ such that $VCD(\mathcal{C}) = d$.

Let $K$ be a set of cardinality $m$ and let $k_0$ be a fixed (but, otherwise, arbitrary) element of $K$. Consider the trace $\mathcal{C}_{K-\{k_0\}}$. Since $|K-\{k_0\}| = m-1$, we have, by the inductive hypothesis, $|\mathcal{C}_{K-\{k_0\}}| \leq \phi(d, m-1)$.

Let $\mathcal{C}'$ be the collection of sets given by

$$\mathcal{C}' = \{G \in \mathcal{C}_K \mid k_0 \notin G, G \cup \{k_0\} \in \mathcal{C}_K\}.$$

Observe that $\mathcal{C}' = \mathcal{C}'_{K-\{k_0\}}$ because $\mathcal{C}'$ consists only of subsets of $K - \{k_0\}$. Further, note that the Vapnik-Chervonenkis dimension of $\mathcal{C}'$ is less than $d$. Indeed, let $K'$ be a subset of $K - \{k_0\}$ that is shattered by $\mathcal{C}'$. Then, $K' \cup \{k_0\}$ is shattered by $\mathcal{C}$. hence $|K'| < d$. By the inductive hypothesis, $|\mathcal{C}'| = |\mathcal{C}_{K-\{k_0\}}| \leq \phi(d-1, m-1)$.

The collection of sets $\mathcal{C}_K$ is a collection of subsets of $K$ that can be regarded as the union of two disjoint collections: those subsets in $\mathcal{C}_K$ that do not contain the element $k_0$. and those subsets of $K$ that contain $k_0$. The first type of subsets forms the collection $\mathcal{C}_{K-\{k_0\}}$. If $L$ is a second type of subset, then $L - \{k_0\}$ is clearly a member of $\mathcal{C}'$. Thus, we have

$$|\mathcal{C}_K| = |\mathcal{C}_{K-\{k_0\}}| + |\mathcal{C}'_{K-\{k_0\}}|$$

This equality implies

$$|\mathcal{C}_K| \leq \phi(d, m-1) + \phi(d-1, m-1),$$

which is the desired conclusion. □

**Lemma 15.11.** *For $d \in \mathbb{N}$ and $d \geq 2$ we have*

$$2^{d-1} \leq \frac{d^d}{d!}.$$

*Proof.* The argument is by induction on $d$. In the basis step, $d = 2$ both members are equal to 2.

Suppose the inequality holds for $d$. We have

$$\begin{aligned}\frac{(d+1)^{d+1}}{(d+1)!} &= \frac{(d+1)^d}{d!} \\ &= \frac{d^d}{d!} \cdot \frac{(d+1)^d}{d^d} \\ &= \frac{d^d}{d!} \cdot \left(1+\frac{1}{d}\right)^d \\ &\geq 2^d \cdot \left(1+\frac{1}{d}\right)^d \\ &\quad \text{(by inductive hypothesis)} \\ &\geq 2^d\end{aligned}$$

because

$$\left(1+\frac{1}{d}\right)^d \geq 1 + d\frac{1}{d} = 2.$$

This concludes the proof of the inequality. □

**Lemma 15.12.** *The function $\phi$ satisfies the inequality:*

$$\phi(d,m) \leq 2\frac{m^d}{d!}$$

*for every $m \geq d$ and $d \geq 1$.*

*Proof.* The argument is by induction on $d$ and $n$. If $d = 1$, then $\phi(1,m) = m+1 \leq 2m$ for $m \geq 1$, so the inequality holds for every $m \geq 1$, when $d = 1$.

If $m = d \geq 2$, then $\phi(d,m) = \phi(d,d) = 2^d$ and the desired inequality follows immediately from Lemma 15.11.

Suppose that the inequality holds for $m > d \geq 1$. We have

$$\begin{aligned}\phi(d,m+1) &= \phi(d,m) + \phi(d-1,m) \\ &\quad \text{(by the definition of } \phi) \\ &\leq 2\frac{m^d}{d!} + 2\frac{m^{d-1}}{(d-1)!} \\ &\quad \text{(by inductive hypothesis)} \\ &= 2\frac{m^{d-1}}{(d-1)!}\left(1+\frac{m}{d}\right).\end{aligned}$$

It is easy to see that the inequality

$$2\frac{m^{d-1}}{(d-1)!}\left(1+\frac{m}{d}\right) \leq 2\frac{(m+1)^d}{d!}$$

is equivalent to

$$\frac{d}{m} + 1 \leq \left(1 + \frac{1}{m}\right)^d$$

and, therefore, is valid. This yields immediately the inequality of the lemma. □

The next theorem discusses the asymptotic behavior of the function $\phi$:

**Theorem 15.13.** *The function $\phi$ satisfies the inequality:*

$$\phi(d, m) < \left(\frac{em}{d}\right)^d$$

*for every $m \geq d$ and $d \geq 1$.*

*Proof.* From Lemma 15.12 we know that $\phi(d, m) \leq 2\frac{m^d}{d!}$. Therefore, we need to show only that

$$2\left(\frac{d}{e}\right)^d < d!.$$

The argument is by induction on $d \geq 1$. The basis case, $d = 1$ is immediate. Suppose that $2\left(\frac{d}{e}\right)^d < d!$. We have

$$\begin{aligned}
&2\left(\frac{d+1}{e}\right)^{d+1} \\
&= 2\left(\frac{d}{e}\right)^d \left(\frac{d+1}{d}\right)^d \frac{d+1}{e} \\
&= \left(1 + \frac{1}{d}\right)^d \frac{1}{e} \cdot 2\left(\frac{d}{e}\right)^d (d+1) \\
&< 2\left(\frac{d}{e}\right)^d (d+1),
\end{aligned}$$

because

$$\left(1 + \frac{1}{d}\right)^d < e.$$

The last inequality holds because the sequence $\left(\left(1 + \frac{1}{d}\right)^d\right)_{d \in \mathbb{N}}$ is an increasing sequence whose limit is $e$. Since $2\left(\frac{d+1}{e}\right)^{d+1} < 2\left(\frac{d}{e}\right)^d (d+1)$, by inductive hypothesis we obtain:

$$2\left(\frac{d+1}{e}\right)^{d+1} < (d+1)!.$$

This proves the inequality of the theorem. □

**Corollary 15.14.** *If $m$ is sufficiently large we have $\phi(d, m) = O(m^d)$.*

*Proof.* The statement is a direct consequence of Theorem 15.13. □

Let $u : B_2^k \longrightarrow B_2$ be a Boolean function of $k$ arguments and let $C_1, \ldots, C_k$ be $k$ subsets of a set $U$. Define the set $u(C_1, \ldots, C_k)$ as the subset $C$ of $U$ whose indicator function is $I_C = u(I_{C_1}, \ldots, I_{C_k})$.

*Example 15.15.* If $u : B_2^2 \longrightarrow B_2$ is the Boolean function $u(a_1, a_2) = a_1 \vee a_2$, then $u(C_1, C_2)$ is $C_1 \cup C_2$; similarly, if $u(x_1, x_2) = x_1 \oplus x_2$, then $u(C_1, C_2)$ is the symmetric difference $C_1 \oplus C_2$ for every $C_1, C_2 \in \mathcal{P}(U)$.

Let $u : B_2^k \longrightarrow B_2$ and $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are $k$ family of subsets of $U$, the family of sets $u(\mathcal{C}_1, \ldots, \mathcal{C}_k)$ is

$$u(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \{u(C_1, \ldots, C_k) \mid C_1 \in \mathcal{C}_1, \ldots, C_k \in \mathcal{C}_k\}.$$

**Theorem 15.16.** *Let $\alpha(k)$ be the least integer $a$ such that $\frac{a}{\log(ea)} > k$.*

*If $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are $k$ collections of subsets of the set $U$ such that $d = \max\{VCD(\mathcal{C}_i) \mid 1 \leq i \leq k\}$ and $u : B_2^2 \longrightarrow B_2$ is a Boolean function, then*

$$VCD(u(\mathcal{C}_1, \ldots, \mathcal{C}_k)) \leq \alpha(k) \cdot d.$$

*Proof.* Let $S$ be a subset of $U$ that consists of $m$ elements. The collection $(\mathcal{C}_i)_S$ is not larger than $\phi(d, m)$. For a set in the collection $W \in u(\mathcal{C}_1, \ldots, \mathcal{C}_k)_S$ we can write $W = S \cap u(C_1, \ldots, C_k)$, or, equivalently, $1_W = 1_S \cdot u(1_{C_1}, \ldots, 1_{C_k})$. By Exercise 20 of Chapter 5, there exists a Boolean function $g_S$ such that

$$1_S \cdot u(1_{C_1}, \ldots, 1_{C_k}) = g_S(1_S \cdot 1_{C_1}, \ldots, 1_S \cdot 1_{C_k}) = g_S(1_{S \cap C_1}, \ldots, 1_{S \cap C_k}).$$

Since there are at most $\phi(d, m)$ distinct sets of the form $S \cap C_i$ for every $i$, $1 \leq i \leq k$, it follows that there are at most $(\phi(d, m))^k$ distinct sets $W$, hence $u(\mathcal{C}_1, \ldots, \mathcal{C}_k)[m] \leq (\phi(d, m))^k$.

Theorem 15.13 implies

$$u(\mathcal{C}_1, \ldots, \mathcal{C}_k)[m] \leq \left(\frac{em}{d}\right)^{kd}.$$

We observed that if $\mathcal{C}[m] < 2^m$, then $VCD(\mathcal{C}) < m$. Therefore, to limit the Vapnik-Chervonenkis dimension of the collection $u(\mathcal{C}_1, \ldots, \mathcal{C}_k)$ it suffices to require that $\left(\frac{em}{d}\right)^{kd} < 2^m$.

Let $a = \frac{m}{d}$. The last inequality can be written as $(ea)^{kd} < 2^{ad}$; equivalently, we have $(ea)^k < 2^a$, which yields $k < \frac{a}{\log(ea)}$. If $\alpha(k)$ is the least integer $a$ such that $k < \frac{a}{\log(ea)}$, then $m \leq \alpha(k)d$, which gives our conclusion. □

*Example 15.17.* If $k = 2$, the least integer $a$ such that $\frac{a}{\log(ea)} > 2$ is $k = 10$, as it can be seen by graphing this function; thus, if $\mathcal{C}_1, \mathcal{C}_2$ are two collection of concepts with $VCD(\mathcal{C}_1) = VCD(\mathcal{C}_2) = d$, the Vapnik-Chervonenkis dimension of the collections $\mathcal{C}_1 \vee \mathcal{C}_2$ or $\mathcal{C}_1 \wedge \mathcal{C}_2$ is not larger than $10d$.

**Lemma 15.18.** *Let $S, T$ be two sets and let $f : S \longrightarrow T$ be a function. If $\mathcal{D}$ is a collection of subsets of $T$, $U$ is a finite subset of $S$ and $\mathcal{C} = f^{-1}(\mathcal{D})$ is the collection $\{f^{-1}(D) \mid D \in \mathcal{D}\}$, then $|\mathcal{C}_U| \leq |\mathcal{D}_{f(U)}|$.*

*Proof.* Let $V = f(U)$ and denote $f \restriction_U$ by $g$. For $D, D' \in \mathcal{D}$ we have

$$\begin{aligned}&(U \cap f^{-1}(D)) \oplus (U \cap f^{-1}(D'))\\ &= U \cap (f^{-1}(D) \oplus f^{-1}(D')) = U \cap (f^{-1}(D \oplus D'))\\ &= g^{-1}(V \cap (D \oplus D')) = g^{-1}(V \cap D) \oplus g^{-1}(V \oplus D').\end{aligned}$$

Thus, $C = U \cap f^{-1}(D)$ and $C' = U \cap f^{-1}(D')$ are two distinct members of $\mathcal{C}_U$, then $V \cap D$ and $V \cap D'$ are two distinct members of $\mathcal{D}_{f(U)}$. This implies $|\mathcal{C}_U| \leq |\mathcal{D}_{f(U)}|$. □

**Theorem 15.19.** *Let $S, T$ be two sets and let $f : S \longrightarrow T$ be a function. If $\mathcal{D}$ is a collection of subsets of $T$ and $\mathcal{C} = f^{-1}(\mathcal{D})$ is the collection $\{f^{-1}(D) \mid D \in \mathcal{D}\}$, then $VCD(\mathcal{C}) \leq VCD(\mathcal{D})$. Moreover, if $f$ is a surjection, then $VCD(\mathcal{C}) = VCD(\mathcal{D})$.*

*Proof.* Suppose that $\mathcal{C}$ shatters an $n$-element subset $K = \{x_1, \ldots, x_n\}$ of $S$, so $|\mathcal{C}_K| = 2^n$ By Lemma 15.18 we have $|\mathcal{C}_K| \leq |\mathcal{D}_{f(U)}|$, so $|\mathcal{D}_{f(U)}| \geq 2^n$, which implies $|f(U)| = n$ and $|\mathcal{D}_{f(U)}| = 2^n$, because $f(U)$ cannot have more than $n$ elements. Thus, $\mathcal{D}$ shatters $f(U)$, so $VCD(\mathcal{C}) \leq VCD(\mathcal{C})$.

Suppose now that $f$ is surjective and $H = \{t_1, \ldots, t_m\}$ is an $m$ element set that is shattered by $\mathcal{D}$. Consider the set $L = \{u_1, \ldots, u_m\}$ such that $u_i \in f^{-1}(t_i)$ for $1 \leq i \leq m$. Let $U$ be a subset of $L$. Since $H$ is shattered by $\mathcal{D}$, there is a set $D \in \mathcal{D}$ such that $f(U) = H \cap D$, which implies $U = L \cap f^{-1}(D)$. Thus, $L$ is shattered by $\mathcal{C}$ and this means that $VCD(\mathcal{C}) = VCD(\mathcal{D})$. □

**Definition 15.20.** *The* density *of $\mathcal{C}$ is the number*

$$\mathit{dens}(\mathcal{C}) = \inf\{s \in \mathbb{R}_{>0} \mid \mathcal{C}[m] \leq c \cdot m^s \text{ for every } m \in \mathbb{N}\},$$

*for some positive constant $c$.*

**Theorem 15.21.** *Let $S, T$ be two sets and let $f : S \longrightarrow T$ be a function. If $\mathcal{D}$ is a collection of subsets of $T$ and $\mathcal{C} = f^{-1}(\mathcal{D})$ is the collection $\{f^{-1}(D) \mid D \in \mathcal{D}\}$, then $\mathit{dens}(\mathcal{C}) \leq \mathit{dens}(\mathcal{D})$. Moreover, if $f$ is a surjection, then $\mathit{dens}(\mathcal{C}) = \mathit{dens}(\mathcal{D})$.*

*Proof.* Let $L$ be a subset of $S$ such that $|L| = m$. Then, $|\mathcal{C}_L| \leq |\mathcal{D}_{f(L)}|$. In general, we have $|f(L)| \leq m$, so $|\mathcal{D}_{f(L)}| \leq \mathcal{D}[m] \leq cm^s$. Therefore, by Lemma 15.18, we have $|\mathcal{C}_L| \leq |\mathcal{D}_{f(L)}| \leq \mathcal{D}[m] \leq cm^s$, which implies $\mathsf{dens}(\mathcal{C}) \leq \mathsf{dens}(\mathcal{D})$.

If $f$ is a surjection, then, for every finite subset $M$ of $T$ such that $|M| = m$ there is a subset $L$ of $S$ such that $|L| = |M|$ and $f(L) = M$. Therefore, $\mathcal{D}[m] \leq \mathcal{C}[m]$ and this implies $\mathsf{dens}(\mathcal{C}) = \mathsf{dens}(\mathcal{D})$. □

If $\mathcal{C}, \mathcal{D}$ are two collections of sets such that $\mathcal{C} \subseteq \mathcal{D}$, then $VCD(\mathcal{C}) \leq VCD(\mathcal{D})$ and $\mathsf{dens}(\mathcal{C}) \leq \mathsf{dens}(\mathcal{D})$.

**Theorem 15.22.** *Let $\mathcal{C}$ be a collection of subsets of a set $S$ and let $\mathcal{C}' = \{S - C \mid C \in \mathcal{C}\}$. Then, for every $K \in \mathcal{P}(S)$ we have $|\mathcal{C}_K| = |\mathcal{C}'_K|$.*

*Proof.* We will prove the statement by showing the existence of a bijection $f : \mathcal{C}_K \longrightarrow \mathcal{C}'_K$. If $U \in \mathcal{C}_K$, then $U = K \cap C$, where $C \in \mathcal{C}$. Then $S - C \in \mathcal{C}'$ and we define $f(U) = K \cap (S - C) = K - C \in \mathcal{C}'_K$. Note that $f$ is well-defined because if $K \cap C_1 = K \cap C_2$, then $K - C_1 = K - (K \cap C_1) = K - (K \cap C_2) = K - C_2$.

It is clear that if $f(U) = f(V)$ for $U, V \in \mathcal{C}_K$, $U = K \cap C_1$, and $V = K \cap C_2$, then $K - C_1 = K - C_2$, so $K \cap C_1 = K \cap C_2$ and this means that $U = V$. Thus, $f$ is injective. If $W \in \mathcal{C}'_K$, then $W = K \cap C'$ for some $C' \in \mathcal{C}$. Since $C' = S - C$ for some $C \in \mathcal{C}$, it follows that $W = K - C$, so $W = f(U)$, where $U = K \cap C$. □

**Corollary 15.23.** *Let $\mathcal{C}$ be a collection of subsets of a set $S$ and let $\mathcal{C}' = \{S - C \mid C \in \mathcal{C}\}$. We have $\mathsf{dens}(\mathcal{C}) = \mathsf{dens}(\mathcal{C}')$ and $VCD(\mathcal{C}) = VCD(\mathcal{C}')$.*

*Proof.* This statement follows immediately from Theorem 15.22. □

**Theorem 15.24.** *For every collection of sets we have $\mathsf{dens}(\mathcal{C}) \leq VCD(\mathcal{C})$. Furthermore, if $\mathsf{dens}(\mathcal{C})$ is finite, then $\mathcal{C}$ is a VC-class.*

*Proof.* If $\mathcal{C}$ is not a VC-class the inequality $\mathsf{dens}(\mathcal{C}) \leq VCD(\mathcal{C})$ is clearly satisfied. Suppose now that $\mathcal{C}$ is a VC-class and $VCD(\mathcal{C}) = d$. By Sauer-Shelah Theorem (Theorem 15.10) we have $\mathcal{C}[m] \leq \phi(d, m)$; then, by Theorem 15.13, we obtain $\mathcal{C}[m] \leq \left(\frac{em}{d}\right)^d$, so $\mathsf{dens}(\mathcal{C}) \leq d$.

Suppose now that $\mathsf{dens}(\mathcal{C})$ is finite. Since $\mathcal{C}[m] \leq cm^s \leq 2^m$ for $m$ sufficiently large, it follows that $VCD(\mathcal{C})$ is finite, so $\mathcal{C}$ is a VC-class. □

Let $\mathcal{D}$ be a finite collection of subsets of a set $S$. In Supplement 6 of Chapter 1 the partition $\pi_{\mathcal{D}}$ was defined as consisting of the nonempty sets of the form $\{D_1^{a_1} \cap D_2^{a_2} \cap \cdots \cap D_r^{a_r}$, where $(a_1, a_2, \ldots, a_r) \in \{0, 1\}^r$.

**Definition 15.25.** *A collection $\mathcal{D} = \{D_1, \ldots, D_r\}$ of subsets of a set $S$ is* independent *if the partition $\pi_{\mathcal{D}}$ has the maximum numbers of blocks, that is, it consists of $2^r$ blocks.*

If $\mathcal{D}$ is independent, then the Boolean subalgebra generated by $\mathcal{D}$ in the Boolean algebra $(\mathcal{P}(S), \{\cap, \cup, {}^-, \emptyset, S\})$ contains $2^{2^r}$ sets, because this subalgebra has $2^r$ atoms. Thus, if $\mathcal{D}$ shatters a subset $T$ with $|T| = p$, then the collection $\mathcal{D}_T$ contains $2^p$ sets, which implies $2^p \leq 2^{2^r}$, or $p \leq 2^r$.

Let $\mathcal{C}$ be a collection of subsets of a set $S$. The *independence number of* $\mathcal{C}$, $I(\mathcal{C})$ is:

$$I(\mathcal{C}) = \sup\{r \mid \{C_1, \ldots, C_r\} \text{ is independent for some finite } \{C_1, \ldots, C_r\} \subseteq \mathcal{C}\}.$$

The next theorem is an analog of Theorem 15.19 for the independence number of a collection.

**Theorem 15.26.** *Let $S, T$ be two sets and let $f : S \longrightarrow T$ be a function. If $\mathcal{D}$ is a collection of subsets of $T$ and $\mathcal{C} = f^{-1}(\mathcal{D})$ is the collection $\{f^{-1}(D) \mid D \in \mathcal{D}\}$, then $I(\mathcal{C}) \leq I(\mathcal{D})$. Moreover, if $f$ is a surjection, then $I(\mathcal{C}) = I(\mathcal{D})$.*

*Proof.* Let $\mathcal{E} = \{D_1, \ldots, D_p\}$ be an independent finite subcollection of $\mathcal{D}$. The partition $\pi_{\mathcal{E}}$ contains $2^r$ blocks. By Supplement 23 of Chapter 5, the number of atoms of the subalgebra generated by $\{f^{-1}(D_1), \ldots, f^{-1}(D_p)\}$ is not greater than $2^r$. Therefore, $I(\mathcal{C}) \leq I(\mathcal{D})$; from the same supplement it follows that if $f$ is surjective, then $I(\mathcal{C}) = I(\mathcal{D})$. □

**Theorem 15.27.** *If $\mathcal{C}$ is a collection of subsets of a set $S$ such that $VCD(\mathcal{C}) \geq 2^n$, then $I(\mathcal{C}) \geq n$.*

*Proof.* Suppose that $VCD(\mathcal{C}) \geq 2^n$, that is, there exists a subset $T$ of $S$ that is shattered by $\mathcal{C}$ and has at least $2^n$ elements. Then, the collection $\mathcal{C}_T$ contains at least $2^{2^n}$ sets, which means that the Boolean subalgebra of $\mathcal{P}(T)$ generated by $\mathcal{T}_C$ contains at least $2^n$ atoms. This implies that the subalgebra of $\mathcal{P}(S)$ generated by $\mathcal{C}$ contains at least this number of atoms, so $I(\mathcal{C}) \geq n$. □

## 15.3 Perceptrons

**Definition 15.28.** *Let $\boldsymbol{w} \in \mathbb{R}^n$ be a $n$-dimensional vector, and let $t \in \mathbb{R}$ be a number.*

*A* perceptron *is a collection of functions $\mathfrak{P}^n = \{P^n_{\boldsymbol{w},t} \mid \boldsymbol{w} \in \mathbb{R}^n, t \in \mathbb{R}\}$, where a function $P^n_{\boldsymbol{w},t} : \mathbb{R}^n \longrightarrow \{0, 1\}$ is defined by*

$$f_{\boldsymbol{w},t}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{w}\boldsymbol{x} \geq t, \\ 0 & \text{otherwise,} \end{cases}$$

*for $\boldsymbol{x} \in \mathbb{R}^n$.*

*We refer to $w$ as the* weight vector *and to $t$ as the* threshold *of the function $P^n_{\boldsymbol{w},t}$. The set $\mathbb{R}^{n+1}$ of all pairs $(\boldsymbol{w}, t)$ is the* parameter space *of the perceptron.*

Define the function $sign : \mathbb{R} \longrightarrow \{0, 1\}$ by

$$sign(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

for $x \in \mathbb{R}$. Now, the function $P^n_{\mathbf{w},t}$ can be written simply as $P^n_{\mathbf{w},t}(\mathbf{x}) = sign(\mathbf{wx} - t)$ for $\mathbf{x} \in \mathbb{R}^n$.

Each function $P^n_{\mathbf{w},t}$ generates a hyperplane $H_{\mathbf{w},t}$ in $\mathbb{R}^n$ given by

$$H_{\mathbf{w},t} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{wx} - t = 0\}.$$

The set $R^n - H_{\mathbf{w},t}$ has two connected components,

$$H^+_{\mathbf{w},t} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{wx} - t > 0\},$$
$$H^-_{\mathbf{w},t} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{wx} - t < 0\}.$$

which are both half-spaces of $\mathbb{R}^n$ and, also are convex sets. Furthermore, $H^+_{\mathbf{w},t}$ and $H^-_{\mathbf{w},t}$ are clearly disjoint.

The next statement is a generalization of the computation of Example 15.7.

**Theorem 15.29.** *Let $\mathfrak{P}^n$ be the perceptron with $n$ inputs. For the collection $\mathcal{S}$ of half-spaces generated by the perceptron we have $VCD(\mathcal{S}) = n + 1$.*

*Proof.* We show first that no subset $S$ of $\mathbb{R}^n$ that consists of $n+2$ points can be shattered by $\mathcal{S}$.

Indeed, suppose that $\mathcal{S}$ would shatter a set $S$ with $|S| = n+2$. By Radon's Theorem (Theorem B.13) there are two disjoint subsets $R, Q$ of $S$ such that $S = R \cup Q$ and $\mathbf{K}_{conv}(R) \cap \mathbf{K}_{conv}(Q) \neq \emptyset$. Suppose that there is a half-space $H^+_{\mathbf{w},t}$ such that $R \subseteq H^+_{\mathbf{w},t}$ and $Q \subseteq H^-_{\mathbf{w},t}$. This would imply $\mathbf{K}_{conv}(R) \subseteq H^+_{\mathbf{w},t}$ and $\mathbf{K}_{conv}(Q) \subseteq H^-_{\mathbf{w},t}$. In turn, this implies

$$\mathbf{K}_{conv}(R) \bigcap \mathbf{K}_{conv}(Q) \subseteq H^+_{\mathbf{w},t} \cap H^-_{\mathbf{w},t} = \emptyset,$$

which contradicts Radon's theorem. Consequently, $VCD(\mathcal{S}) \leq n + 1$.

To prove the converse inequality let $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^n$, where 1 occurs in the $i$th place for $1 \leq i \leq n$. We will prove that the collection $\mathcal{S}$ shatters the set $T = \{\mathbf{0}\} \cup \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$.

Let $U$ be a subset of $T$. Define $\mathbf{w} = (w_1, \ldots, w_n)$ by $w_i = 2I_U(\mathbf{e}_i) - 1$ for $1 \leq i \leq n$, where $I_U : \mathcal{P}(T) \longrightarrow \{0, 1\}$ is the indicator function of $U$. Also, define $t$ as

$$t = \frac{1 - 2 \cdot I_U(\mathbf{0})}{2}$$

We claim that $T \cap H^+_{\mathbf{w},t} = U$. Indeed, we have $\mathbf{x} \in T \cap H^+_{\mathbf{w},t}$ if and only if $\mathbf{wx} > t$.

To prove the inclusion $T \cap H^+_{\mathbf{w},t} \subseteq U$ we need to consider two cases:

(i) If $\mathbf{x} = \mathbf{0} \in T \cap H^+_{\mathbf{w},t}$, then $\mathbf{wx} > t$ implies $t < 0$, which happens if $I_U(\mathbf{0}) = 1$, that is, if $\mathbf{0} \in U$.

(ii) If $\mathbf{x} = \mathbf{e}_i$, then $\mathbf{wx} > t$ is equivalent to $w_i > t$, that is $2I_U(\mathbf{e}_i) - 1 > t$. Since $t \in \{0.5, -0.5\}$, this implies $2I_U(\mathbf{e}_i) > 0.5$, so $I_U(\mathbf{e}_i) > 0$, that is $I_U(\mathbf{e}_i) = 1$, which is equivalent to $\mathbf{e}_i \in U$.

This shows that $T \cap H^+_{\mathbf{w},t} \subseteq U$.

Conversely, if $\mathbf{x} \in U$ and $\mathbf{x} = \mathbf{0}$ we have $t = -0.5$ and, therefore $\mathbf{wx} = 0 > t$, which implies $\mathbf{0} \in T \cap H^+_{\mathbf{w},t}$. If $\mathbf{x} \in U$ and $\mathbf{x} = \mathbf{e}_i$, then $\mathbf{wx} = w_i = 1 > t$, so, again, $\mathbf{x} \in T \cap H^+_{\mathbf{w},t}$.

This shows that $\mathcal{S}$ shatters a set with $n+1$ elements, so $VCD(\mathcal{S}) \geq n+1$. □

**Corollary 15.30.** *For the collection $\mathcal{S}$ of half-spaces generated by the perceptron $\mathfrak{P}^n$ we have $\mathcal{S}[m] \leq \phi(n+1, m)$.*

*Proof.* The statement follows from Theorems 15.29 and 15.10. □

## Exercises and Supplements

1. Let $\mathcal{C}, \mathcal{D}$ be two collections of subsets of a set $S$. Prove that for every $m \in \mathbb{N}$ we have $(\mathcal{C} \cup \mathcal{D})[m] = \max\{\mathcal{C}[m], \mathcal{D}[m]\}$.
2. Let $S$ be a nonempty set and let $\mathcal{C} = \{\{x\} \mid x \in S\}$. Prove that $VCD(\mathcal{C}) = 1$.
3. Let $S$ be a nonempty set. Prove that if $\mathcal{C}$ is a collection of subsets of $S$ such that $|\mathcal{C}| \geq 2$, then $VCD(\mathcal{C}) \geq 1$.
4. Let $U$ be a finite set and let $\mathcal{C}$ be a collection of subsets of $U$ such that $|\mathcal{C}| \geq 2$. Prove that $VCD(\mathcal{C}) > \frac{\ln |\mathcal{C}|}{1+\ln |U|}$.

   **Solution:** Observe that $\mathcal{C}[|U|] = |\mathcal{C}|$. Therefore, by Sauer-Shelah Theorem (Theorem 15.10) and by Theorem 15.13, we have

   $$|\mathcal{C}| \leq \left(\frac{e|U|}{d}\right)^d,$$

   where $d$ is the VC dimension of the collection $\mathcal{C}$. The last inequality implies

   $$\ln |\mathcal{C}| \leq d(1 + \ln |U| - \ln d),$$

   so $\ln |\mathcal{C}| \leq d(1 + \ln |U|)$, which gives the desired inequality.
5. Prove that if $\mathcal{C}$ is a chain of subsets of a set $S$, then $VCD(\mathcal{C}) = 1$.
6. Let $\mathcal{C}$ be a collection of subsets of $S$. Prove that if $T$ is a subset of $S$, then $VCD(\mathcal{C}_T) \leq VCD(\mathcal{C})$.
7. Let $\mathcal{C}$ be a collection of sets such that $C, C' \in \mathcal{C}$ and $C \neq C'$ implies $C \cap C' = \emptyset$. Prove that $VCD(\mathcal{C}) = 1$.
8. Let $S$ be a set and let $\mathcal{C}_1, \ldots, \mathcal{C}_n$ be $n$ chains in the poset $(\mathcal{P}(S), \subseteq)$. Define the collection $\mathcal{C}$ as $\mathcal{C} = \{\bigcap_{i=1}^n C_i \mid C_i \in \mathcal{C}_i, 1 \leq i \leq n\}$. Prove that $VCD(\mathcal{C}) \leq n$.

   **Solution:** Let $T$ be a subset of $S$ such that $|T| = n+1$. Clearly, $T$ has $n+1$ subsets that have $n$ elements.

   For each $i$ at most one $n$-element subset of $T$ is the intersection of the form $T \cap C$, where $C \in \mathcal{C}_i$. Indeed, if we would have two distinct $n$-element

sets of the form $T \cap C'$ and $T \cap C''$, where $C', C'' \in \mathcal{C}$ this would imply the existence of $x' \in (T \cap C') - (T \cap C'')$ and of $x'' \in (T \cap C'') - (T \cap C')$, which would mean that $x' \in C' - C''$ and $x'' \in C'' - C'$, thus contradicting the $\mathcal{C}_i$ is a chain of sets. Let $U_i$ be this $n$-element when it exists.

Let $W$ be an $n$-element subset of $T$ such that $W = T \cap C$ for some $C = \bigcap_{i=1}^{n} C_i \in \mathcal{C}$. Then, either $C_j \cap T = W$ or $C_i \cap T = T$ for $1 \leq j \leq n$ and $C_i \cap T = W$ for at least one $i$, $1 \leq i \leq n$. Therefore, $W = U_i$ for some $i$, $1 \leq i \leq n$, which shows that at most $n$ subsets of $T$ that contain $n$ elements can be obtained as intersections of $T$ with the elements of $\mathcal{C}$. Thus, $T$ is not shattered by $\mathcal{C}$ and $VCD(\mathcal{C}) \leq n$.

9. For $1 \leq i \leq n$ and $a \in \mathbb{R}$ let $C_{i,a} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = (x_1, \ldots, x_n), x_i \leq a\}$. The chain of sets $\mathcal{C}_i$ is defined by $\{C_{i,a} \mid a \in \mathbb{R}\}$ for $1 \leq i \leq n$.
Prove that $\mathcal{C} = \bigcap_{i=1}^{n} \mathcal{C}_i$ shatters the set $B = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, where $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ has 1 as its $i$th component for $1 \leq i \leq n$, so $VCD(\mathcal{C}) = n$.
10. The statement included here is a generalization of Example 15.8. Prove that the Vapnik-Chervonenkis dimension of the collection of rectangular subsets of $\mathbb{R}^n$ given by

$$\mathcal{C} = \{\prod_{i=1}^{n} [a_i, b_i] \mid a_i, b_i \in \hat{\mathbb{R}}, a_i \leq b_i, \text{ for } 1 \leq i \leq n\}$$

is $2n$. If $a_i = -\infty$ for all $a_i$, $1 \leq i \leq n$, then $VCD(\mathcal{C}) = n$.
11. Let $S$ be a set that contains at least two elements and let $\mathcal{C}$ be a collection of subsets $S$. Suppose that for every two-element subset of $S$, $T = \{t_1, t_2\}$, there exist $U, V \in \mathcal{C}$ such that $T \subseteq U$ and $T \cap V = \emptyset$. Then $VCD(\mathcal{C}) = 1$ if and only if $\mathcal{C}$ is a chain.

**Solution:** Suppose that $VCD(\mathcal{C}) = 1$ but $\mathcal{C}$ is not a chain. Then, $\mathcal{C}$ contains two sets $C', C''$ such that neither $C' \subseteq C''$ nor $C'' \subseteq C'$. Let $c' \in C' - C''$ and $c'' \in C'' - C'$. Then, the two element set $T = \{c', c''\}$ is shattered by $\mathcal{C}$, which implies $VCD(\mathcal{C}) \geq 2$. The reverse implication follows from Supplement 5.
12. Prove that if $\mathcal{C}$ is a collection of subsets of a set $S$ such that $\{\emptyset, S\} \subseteq \mathcal{C}$, then $VCD(\mathcal{C}) = 1$ if and only if $\mathcal{C}$ is a chain.
13. Let $S$ be a finite set and let $\mathcal{C}$ be a collection of subsets of $S$. Prove that $|\mathcal{C}| = |\{K \in \mathcal{P}(S) \mid \mathcal{C} \text{ shatters } K\}|$.

**Solution:** Let $x$ be an element of $S$ and let $\phi_x : \mathcal{C} \longrightarrow \mathcal{P}$ be the injective mapping introduced in Supplement 26 of Chapter 1. We claim that if $\phi_x(\mathcal{C}) = \{\phi_x(C) \mid C \in \mathcal{C}\}$ shatters $K$, then $\mathcal{C}$ shatters $K$. If $x \notin K$, then $\mathcal{C}_K = \phi_x(\mathcal{C})_K$, so the statement obviously holds. If $x \in K$ and $L \subseteq K - \{x\}$, then there is $F \in \phi_x(C)$ such that $F \cap K = L \cup \{x\}$ and $T = \phi_x(C)$ for some $C \in \mathcal{C}$. Since $x \in F$, both $F$ and $F - \{x\}$ belong to $\mathcal{C}$, so $\mathcal{C}$ shatters $K$.

Define $w(\mathcal{C}) = \sum\{|C| \mid C \in \mathcal{C}\}$. Let $\mathcal{C}'$ be a collection of sets obtained from $\mathcal{C}$ by applying transforms of the form $\phi_x$, such that $w(\mathcal{C}')$ is

minimal. For $C \in \mathcal{C}'$ and $x \in K$ we must have $C - \{x\} \in \mathcal{C}'$ because otherwise $w(\phi_x(\mathcal{C}')) < w(\mathcal{C}')$, contradicting the minimality of $\mathcal{C}'$. Thus, $\mathcal{C}'$ is hereditary, so it shatters any set it contains. Since $|\mathcal{C}'| = |\mathcal{C}|$ (by Supplement 26 of Chapter 1, and $\mathcal{C}$ shatters at least as many sets as $\mathcal{C}$ we obtain the desired equality.

## Bibliographical Comments

Theorem 15.16 appears in [142]. Supplements 4-12 contain results obtained in [144]. Note that in [144] the Vapnik-Chervonenkis dimension of a collection of set is defined as the smallest $n$ such that no $n$-element set is shattered by $\mathcal{C}$, so values of $VCD(\mathcal{C})$ in [144] are obtained by increasing by one the value of the $VCD$ adopted here (and in the vast majority of publications).

The notion of density of a collection of sets was introduced by P. Assouad in [6]. Supplement 13 originates in [86].

# Part V

# Appendices

# A

# Asymptotics

We present some formal concepts that allow the evaluation of the rate of growth of algorithm complexity.

**Definition A.1.** *Let $f : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$ be a function. The classes of functions $O(f), \Theta(f)$, and $\Omega(f)$ are given by:*

(i) $0(f)$ *consists of those functions $g : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$ for which there exists $c \in \mathbb{R}_{>0}$ and $n_c \in \mathbb{N}$ such that $n \geq n_c$ implies $g(n) \leq cf(n)$.*

(ii) $\Theta(f)$ *consists of those functions $g : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$ for which there exist $c, c' \in \mathbb{R}_{>0}$ and $n_{c,c'} \in \mathbb{N}$ such that $n \geq n_{c,c'}$ implies $c'f(n) \leq g(n) \leq cf(n)$.*

(iii) $\Omega(f)$ *consists of those functions $g : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$ for which there exists $c \in \mathbb{R}_{>0}$ and $n_c \in \mathbb{N}$ such that $n \geq n_c$ implies $g(n) \geq cf(n)$.*

Another collection of classes of functions is introduced next.

**Definition A.2.** *Let $f : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$ be a function. The classes of functions $o(f), \theta(f)$ and $\omega(f)$ are given by:*

(i) $o(f)$ *consists of those functions $g : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$ for which*

$$\lim_{n \to \infty} \frac{g(n)}{f(n)} = 0.$$

(ii) $\theta(f)$ *consists of those functions $g : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$ for which*

$$\lim_{n \to \infty} \frac{g(n)}{f(n)} = k$$

*for some $k \in \mathbb{R}_{>0}$.*

(iii) $\omega(f)$ *consists of those functions $g : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$ such that*

$$\lim_{n \to \infty} \frac{g(n)}{f(n)} = \infty.$$

**Theorem A.3.** *If $g, h \in R(f)$, where $R \in \{0, \Theta, \Omega, o, \theta, \omega\}$, then $ag + bh \in R(f)$ for every $a, b > 0$.*

*Proof.* The argument is elementary and is left to the reader. □

**Theorem A.4.** *For every function $f : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$, we have the inclusions $o(f) \subseteq O(f)$, $\theta(f) \subseteq \Theta(f)$, and $\omega(f) \subseteq \Omega(f)$.*

*Proof.* We show only the inclusion $\theta(f) \subseteq \Theta(f)$ and leave the two other for the reader.

Suppose that $g \in \theta(f)$; that is, $\lim_{n\to\infty} \frac{g(n)}{f(n)} = k$ for some $k > 0$. By the definition of the limit, for every $\epsilon > 0$ there is $n_\epsilon$ such that $n \geq n_\epsilon$ implies

$$\left| \frac{g(n)}{f(n)} - k \right| < \epsilon$$

or, equivalently,

$$k - \epsilon < \frac{g(n)}{f(n)} < k + \epsilon.$$

Thus, the role of $c, c'$ can be played by $k + \epsilon$ and $k - \epsilon$, respectively, and we may conclude that $g \in \Theta(f)$. □

The reverse inclusions are not necessarily true. For example, if $f$ is a function that differs from the constant function 0 and $g(n) = f(n) \sin \pi n$, then $g \in \Theta(f)$; however, $g \notin \theta(f)$ because $\lim_{n\to\infty} \frac{g(n)}{f(n)} = \lim_{n\to\infty} \sin \pi n$ does not exist.

# B

# Convex Sets and Functions

**Definition B.1.** *Let $(L, +, \cdot)$ be a real linear space and let $C$ be a subset of $L$. The set $C$ is convex if, for all $\boldsymbol{x}, \boldsymbol{y} \in C$ and all $a \in [0, 1]$, we have $(1-a)\boldsymbol{x} + a\boldsymbol{y} \in C$. In other words, every point on the line segment connecting $\boldsymbol{x}$ and $\boldsymbol{y}$ belongs to $C$.*

*Example B.2.* The convex subsets of $(\mathbb{R}, +, \cdot)$ are the intervals of $\mathbb{R}$. Regular polygons are convex subsets of $\mathbb{R}^2$.

**Definition B.3.** *Let $U$ be a subset of a real linear space $(L, +, \cdot)$.*

*A convex combination of $U$ is an element of $L$ of the form $a_1\boldsymbol{x}_1 + \cdots + a_k\boldsymbol{x}_k$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in U$, $a_i \geq 0$ for $1 \leq i \leq k$, and $a_1 + \cdots + a_k = 1$.*

*If the conditions $a_i \geq 0$ are dropped, we have an affine combination of $U$. In other words, $\boldsymbol{x}$ is an affine combination of $U$ if there exist $a_1, \ldots, a_k \in \mathbb{R}$ such that $\boldsymbol{x} = a_1\boldsymbol{x}_1 + \cdots + a_k\boldsymbol{x}_k$, for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in U$, and $\sum_{i=1}^{k} a_i = 1$.*

**Definition B.4.** *Let $U$ be a subset of a real linear space $(L, +, \cdot)$. A subset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is affinely dependent if $\boldsymbol{0} = a_1\boldsymbol{x}_1 + \cdots + a_n\boldsymbol{x}_n$ such that at least one of the numbers $a_1, \ldots, a_n$ is nonzero and $\sum_{i=1}^{n} a_i = 0$. If no such affine combination exists, then $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are affinely independent.*

**Theorem B.5.** *The set $U = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is affinely independent if and only if the set $V = \{\boldsymbol{x}_1 - \boldsymbol{x}_n, \boldsymbol{x}_{n-1} - \boldsymbol{x}_n\}$ is linearly independent.*

*Proof.* Suppose that $U$ is affinely independent but $V$ is linearly dependent; that is, $\mathbf{0} = b_1(\mathbf{x}_1 - \mathbf{x}_n) + \cdots + b_{n-1}(\mathbf{x}_{n-1} - \mathbf{x}_n)$ such that not all numbers $b_i$ are 0. This implies

$$b_1\mathbf{x}_1 + \cdots + b_{n-1}\mathbf{x}_{n-1} - \left(\sum_{i=1}^{n-1} b_i\right)\mathbf{x}_n = \mathbf{0},$$

which contradicts the affine independence of $U$.

Conversely, suppose that $V$ is linearly independent but $U$ is not affinely independent. In this case, $\mathbf{0} = a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n$ such that at least one of the numbers $a_1, \ldots, a_n$ is nonzero and $\sum_{i=1}^{n} a_i = 0$. This implies $a_n = -\sum_{i=1}^{n-1} a_i$, so $\mathbf{0} = a_1(\mathbf{x}_1 - \mathbf{x}_n) + \cdots + a_{n-1}(\mathbf{x}_{n-1} - \mathbf{x}_n)$. Observe that at least one of the numbers $a_1, \ldots, a_{n-1}$ must be distinct from 0 because otherwise we would have $a_1 = \cdots = a_{n-1} = a_n = 0$. This contradicts the linear independence of $V$, so $U$ is affinely independent. □

*Example B.6.* Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two elements of the linear space $(\mathbb{R}^2, +, \cdots)$. The line that passes through $\mathbf{x}_1$ and $\mathbf{x}_2$ consists of all $\mathbf{x}$ such that $\mathbf{x} - \mathbf{x}_1$ and $\mathbf{x} - \mathbf{x}_2$ are collinear; that is, $a(\mathbf{x} - \mathbf{x}_1) + b(\mathbf{x} - \mathbf{x}_2) = 0$ for some $a, b \in \mathbb{R}$ such that $a + b \neq 0$. Thus, we have $\mathbf{x} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2$, where

$$a_1 + a_2 = \frac{a}{a+b} + \frac{b}{a+b} = 1,$$

so $\mathbf{x}$ is an affine combination of $\mathbf{x}_1$ and $\mathbf{x}_2$. It is easy to see that the segment of line contained between $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by a convex combination of $\mathbf{x}_1$ and $\mathbf{x}_2$; that is, by an affine combination $a_1\mathbf{x}_1 + a_2\mathbf{x}_2$ such that $a_1, a_2 \geq 0$.

**Theorem B.7.** *If $C$ is a convex subset of a real linear space $(L, +, \cdot)$, then $C$ contains all convex linear combinations of $C$.*

*Proof.* The proof is by induction on $k \geq 2$ and is left to the reader. □

**Theorem B.8.** *The intersection of any collection of convex sets of a linear space $(L, +, \cdot)$ is a convex set.*

*Proof.* Let $\mathcal{C} = \{C_i \mid i \in I\}$ be a collection of convex sets and let $C = \bigcap \mathcal{C}$. Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_k \in C$, $a_i \geq 0$ for $1 \leq i \leq k$, and $a_1 + \cdots + a_k = 1$. Since $\mathbf{x}_1, \ldots, \mathbf{x}_k \in C_i$, it follows that $a_1\mathbf{x}_1 + \cdots + a_k\mathbf{x}_k \in C_i$ for every $i \in I$. Thus, $a_1\mathbf{x}_1 + \cdots + a_k\mathbf{x}_k \in C$, which proves the convexity of $C$. □

**Corollary B.9.** *The family of convex sets of a linear space $(L, +, \cdots)$ is a closure system on $\mathcal{P}(L)$.*

*Proof.* This statement follows immediately from Theorem B.8 by observing that the set $L$ is convex. □

Corollary B.9 allows us to define the *convex hull* of a subset $U$ of $L$ as the closure $\mathbf{K}_{conv}(U)$ of $U$ relative to the closure system of the convex subsets of $L$. If $U \subseteq \mathbb{R}^n$ consists of $n+1$ points such that no point is an affine combination of the other $n$ points, then $\mathbf{K}_{conv}(U)$ is an *$n$-dimensional simplex in $L$*.

*Example B.10.* A two-dimensional simplex is defined starting from three points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ in $\mathbb{R}^2$ such that none of these points is an affine combination of

the other two (no point is collinear with the others two). Thus, the two-dimensional symplex generated by $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ is the full triangle determined by $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$.

In general, an $n$-dimensional simplex is the convex hull of a set of $n+1$ points $\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}$ in $\mathbb{R}^n$ such that no point is an affine combination of the remaining $n$ points.

Let $S$ be the $n$-dimensional simplex generated by the points $\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}$ in $\mathbb{R}^n$ and let $\mathbf{x} \in S$. If $\mathbf{x} \in S$, then $\mathbf{x}$ is a convex combination of $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{n+1}$. In other words, there exist $a_1, \ldots, a_n, a_{n+1}$ such that $a_1, \ldots, a_n, a_{n+1} \in (0,1)$, $\sum_{i=1}^{n+1} a_i = 1$, and $\mathbf{x} = a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n + a_{n+1}\mathbf{x}_{n+1}$.

The numbers $a_1, \ldots, a_n, a_{n+1}$ are the *baricentric coordinates* of $\mathbf{x}$ relative to the simplex $S$ and are uniquely determined by $\mathbf{x}$. Indeed, if we have

$$\mathbf{x} = a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n + a_{n+1}\mathbf{x}_{n+1} = b_1\mathbf{x}_1 + \cdots + b_n\mathbf{x}_n + b_{n+1}\mathbf{x}_{n+1},$$

and $a_i \neq b_i$ for some $i$, this implies

$$(a_1 - b_1)\mathbf{x}_1 + \cdots + (a_n - b_n)\mathbf{x}_n + (a_{n+1} - b_{n+1})\mathbf{x}_{n+1} = \mathbf{0},$$

which contradicts the affine independence of $\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}$.

The next statement plays a central role in the study of convexity. We reproduce the proof given in [59].

**Theorem B.11 (Carathéodory's Theorem).** *If $U$ is a subset of $\mathbb{R}^n$, then for every $\boldsymbol{x} \in \boldsymbol{K}_{conv}(U)$ we have $\boldsymbol{x} = \sum_{i=1}^{n+1} a_i\boldsymbol{x}_i$, where $\boldsymbol{x}_i \in U$, $a_i \geq 0$ for $1 \leq i \leq n+1$, and $\sum_{i=1}^{n+1} a_i = 1$.*

*Proof.* Consider $\mathbf{x} \in \mathbf{K}_{conv}(U)$. We can write $\mathbf{x} = \sum_{i=1}^{p+1} a_i\mathbf{x}_i$, where $\mathbf{x}_i \in U$, $a_i \geq 0$ for $1 \leq i \leq p+1$, and $\sum_{i=1}^{p+1} a_i = 1$. Let $p$ be the smallest number which allows this kind of expression for $x$. We prove the theorem by showing that $p \leq n$.

Suppose that $p \geq n+1$. Then, the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}\}$ is affinely dependent, so there exist $b_1, \ldots, b_{p+1}$ not all zero such that $\mathbf{0} = \sum_{i=1}^{p+1} b_i\mathbf{x}_i$ and $\sum_{i=1}^{p+1} b_i = 0$. Without loss of generality, we can assume $b_{p+1} > 0$ and $\frac{a_{p+1}}{b_{p+1}} \leq \frac{a_i}{b_i}$ for all $i$ such that $1 \leq i \leq p$ and $b_i > 0$. Define

$$c_i = b_i\left(\frac{a_i}{b_i} - \frac{a_{p+1}}{b_{p+1}}\right)$$

for $1 \leq i \leq p$. We have

$$\sum_{i=1}^{p} c_i = \sum_{i=1}^{p} a_i - \frac{a_{p+1}}{b_{p+1}}\sum_{i=1}^{p} b_i = 1.$$

Furthermore, $c_i \geq 0$ for $1 \leq i \leq p$. Indeed, if $b_i \leq 0$, then $c_i \geq a_i \geq 0$; if $b_i > 0$, then $c_i \geq 0$ because $\frac{a_{p+1}}{b_{p+1}} \leq \frac{a_i}{b_i}$ for all $i$ such that $1 \leq i \leq p$ and $b_i > 0$. Thus, we have

$$\sum_{i=1}^{p} c_i \mathbf{x}_i = \sum_{i=1}^{p} \left( a_i - \frac{a_p}{b_p} b_i \right) \mathbf{x}_i = \sum_{i=1}^{p} a_i \mathbf{x}_i = \mathbf{x},$$

which contradicts the choice of $p$. □

A finite set of points $P$ in $\mathbb{R}^2$ is a *convex polygon* if no member $p$ of $P$ lies in the convex hull of $P - \{p\}$.

**Theorem B.12.** *A finite set of points $P$ in $\mathbb{R}^2$ is a convex polygon if and only if no member $p$ of $P$ lies in a two-dimensional simplex formed by three other members of $P$.*

*Proof.* The argument is straightforward and is left to the reader as an exercise. □

**Theorem B.13 (Radon's Theorem).** *Let $P = \{\boldsymbol{x}_i \in \mathbb{R}^n \mid 1 \leq i \leq n+2\}$ be a set of $n+2$ points in $\mathbb{R}^n$. Then, there are two disjoint subsets $R$ and $Q$ of $P$ such that $\boldsymbol{K}_{conv}(R) \cap \boldsymbol{K}_{conv}(Q) \neq \emptyset$.*

*Proof.* Since $n+2$ points in $\mathbb{R}^n$ are affinely dependent, there exist $a_1, \ldots, a_{n+2}$ not all equal to 0 such that

$$\sum_{i=1}^{n+2} a_i \mathbf{x}_i = \mathbf{0} \tag{B.1}$$

and $\sum_{i=1}^{n+2} a_i = 0$. Without loss of generality, we can assume that the first $k$ numbers are positive and the last $n+2-k$ are not. Let $a = \sum_{i=1}^{k} a_i > 0$ and let $b_j = \frac{a_j}{a}$ for $1 \leq j \leq k$. Similarly, let $c_l = -\frac{a_l}{a}$ for $k+1 \leq l \leq n+2$. Equality (B.1) can now be written as

$$\sum_{j=1}^{k} b_j \mathbf{x}_j = \sum_{l=k+1}^{n+2} c_l \mathbf{x}_l.$$

Since the numbers $b_j$ and $c_l$ are nonnegative and $\sum_{j=1}^{k} b_j = \sum_{l=k+1}^{n+2} c_l = 1$, it follows that $\mathbf{K}_{conv}(\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}) \cap \mathbf{K}_{conv}(\{\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{n+2}\}) \neq \emptyset$. □

**Theorem B.14 (Klein's Theorem).** *If $P \subseteq \mathbb{R}^2$ is a set of five points such that no three of them are collinear, then $P$ contains four points that form a convex quadrilateral.*

*Proof.* Let $P = \{\mathbf{x}_i \mid 1 \leq i \leq\}$. If these five points form a convex polygon, then any four of them form a convex quadrilateral. If exactly one point is in the interior of a convex quadrilateral formed by the remaining four points, then the desired conclusion is reached.

Suppose that none of the previous cases occur. Then, two of the points, say $\mathbf{x}_p, \mathbf{x}_q$, are located inside the triangle formed by the remaining points $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$. Note that the line $\mathbf{x}_p\mathbf{x}_q$ intersects two sides of the triangle $\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k$,

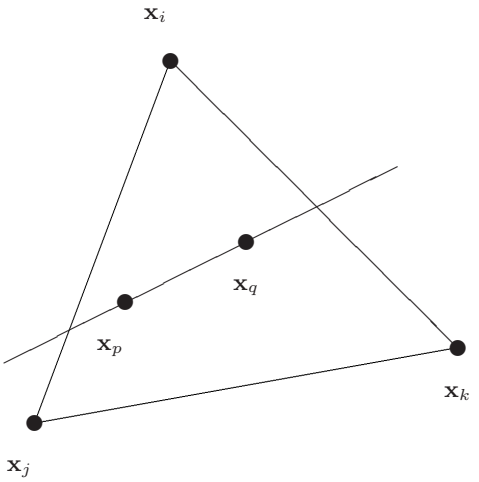



**Fig. B.1.** A five-point configuration in $\mathbb{R}^2$.

say $\mathbf{x}_i\mathbf{x}_j$ and $\mathbf{x}_i\mathbf{x}_k$ (see Figure B.1). Then $\mathbf{x}_p\mathbf{x}_q\mathbf{x}_k\mathbf{x}_j$ is a convex quadrilateral. □

A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is *convex* if its graph on an interval is located below the chord determined by the endpoints of the interval. More formally, we have the following definition.

**Definition B.15.** *A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is convex if $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ for every $x, y \in \mathrm{Dom}(f)$ and $t \in [0,1]$. The function $g : \mathbb{R} \longrightarrow \mathbb{R}$ is concave if $-g$ is convex.*

**Theorem B.16.** *If $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a convex function and $a < b \leq c$, then*

$$\frac{f(b) - f(a)}{b - a} \leq \frac{f(c) - f(a)}{c - a}.$$

*Proof.* Since $a < b \leq c$, we can write $b = ta + (1-t)c$, where $t = \frac{c-b}{c-a} \in (0,1]$. The convexity of $f$ yields the inequality

$$f(b) \leq \frac{c-b}{c-a} f(a) + \frac{b-a}{c-a} f(c),$$

which is easily seen to be equivalent with the desired inequality. □

A similar result follows.

**Theorem B.17.** *If $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a convex function and $a \leq b < c$, then*

$$\frac{f(c) - f(a)}{c - a} \leq \frac{f(c) - f(b)}{c - b}.$$

*Proof.* The argument is similar to the proof of Theorem B.16. □

**Corollary B.18.** *Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a convex function and let $p, q, p', q'$ be four numbers such that $p \leq p' < q \leq q'$. We have the inequality*

$$\frac{f(q) - f(p)}{q - p} \leq \frac{f(q') - f(p')}{q' - p'}. \tag{B.2}$$

*Proof.* By Theorem B.16 applied to the numbers $p', q, q'$, we have

$$\frac{f(q) - f(p')}{q - p'} \leq \frac{f(q') - f(p')}{q' - p'}.$$

Similarly, by applying Theorem B.17 to $p, p', q$, we obtain

$$\frac{f(q) - f(p)}{q - p} \leq \frac{f(q) - f(p')}{q - p'}.$$

The inequality of the corollary can be obtained by combining the last two inequalities. □

From Corollary B.18, it follows that if $f : \mathbb{R} \longrightarrow \mathbb{R}$ is convex and differentiable everywhere, then its derivative is an increasing function.

The converse is also true; namely, if $f$ is differentiable everywhere and its derivative is an increasing function, then $f$ is convex. Indeed, let $a, b, c$ be three numbers such that $a < b < c$. By the mean value theorem, there is $p \in (a, b)$ and $q \in (b, c)$ such that

$$f'(p) = \frac{f(b) - f(a)}{b - a} \text{ and } f'(q) = \frac{f(c) - f(b)}{c - b}.$$

Since $f'(p) \leq f'(q)$, we obtain

$$\frac{f(b) - f(a)}{b - a} \leq \frac{f(c) - f(b)}{c - b},$$

which implies

$$f(b) \leq \frac{c - b}{c - a} f(a) + \frac{b - a}{c - a} f(c);$$

that is, the convexity of $f$. Thus, if $f$ is twice differentiable everywhere and its second derivative is nonnegative everywhere, then it follows that $f$ is convex. Clearly, under the same conditions of differentiability as above, if the second derivative is nonpositive everywhere, then $f$ is concave.

The functions listed in the Table B.1, defined on the set $\mathbb{R}_{\geq 0}$, provide examples of convex (or concave) functions.

**Theorem B.19 (Jensen's Theorem).** *Let $f$ be a function that is convex on an interval $I$. If $t_1, \ldots, t_n \in [0, 1]$ are $n$ numbers such that $\sum_{i=1}^{n} t_i = 1$, then*

$$f\left(\sum_{i=1}^{n} t_i x_i\right) \leq \sum_{i=1}^{n} t_i f(x_i)$$

*for every* $x_1, \ldots, x_n \in I$.

*Proof.* The argument is by induction on $n$, where $n \geq 2$. The basis step, $n = 2$, follows immediately from Definition B.15.

Suppose that the statement holds for $n$, and let $u_1, \ldots, u_n, u_{n+1}$ be $n+1$ numbers such that $\sum_{i=1}^{n+1} u_i = 1$. We have

$$\begin{aligned} &f(u_1 x_1 + \cdots + u_{n-1} x_{n-1} + u_n x_n + u_{n+1} x_{n+1}) \\ &= f\left(u_1 x_1 + \cdots + u_{n-1} x_{n-1} + (u_n + u_{n+1}) \frac{u_n x_n + u_{n+1} x_{n+1}}{u_n + u_{n+1}}\right). \end{aligned}$$

By the inductive hypothesis, we can write

$$\begin{aligned} &f(u_1 x_1 + \cdots + u_{n-1} x_{n-1} + u_n x_n + u_{n+1} x_{n+1}) \\ &\leq u_1 f(x_1) + \cdots + u_{n-1} f(x_{n-1}) + (u_n + u_{n+1}) f\left(\frac{u_n x_n + u_{n+1} x_{n+1}}{u_n + u_{n+1}}\right). \end{aligned}$$

Next, by the convexity of $f$, we have

$$f\left(\frac{u_n x_n + u_{n+1} x_{n+1}}{u_n + u_{n+1}}\right) \leq \frac{u_n}{u_n + u_{n+1}} f(x_n) + \frac{u_{n+1}}{u_n + u_{n+1}} f(x_{n+1}).$$

Combining this inequality with the previous inequality gives the desired conclusion. □

Of course, if $f$ is a concave function and $t_1, \ldots, t_n \in [0, 1]$ are $n$ numbers such that $\sum_{i=1}^{n} t_i = 1$, then

$$f\left(\sum_{i=1}^{n} t_i x_i\right) \geq \sum_{i=1}^{n} t_i f(x_i). \tag{B.3}$$

**Table B.1.** Examples of convex or concave functions.

| Function | Second Derivative | Convexity Property |
|---|---|---|
| $x^r$ for $r > 0$ | $r(r-1)x^{r-2}$ | concave for $r < 1$<br>convex for $r \geq 1$ |
| $\ln x$ | $-\frac{1}{x^2}$ | concave |
| $x \ln x$ | $\frac{1}{x}$ | convex |
| $e^x$ | $e^x$ | convex |

*Example B.20.* We saw that the function $f(x) = \ln x$ is concave. Therefore, if $t_1, \ldots, t_n \in [0,1]$ are $n$ numbers such that $\sum_{i=1}^n t_i = 1$, then

$$\ln\left(\sum_{i=1}^n t_i x_i\right) \geq \sum_{i=1}^n t_i \ln x_i.$$

This inequality can be written as

$$\ln\left(\sum_{i=1}^n t_i x_i\right) \geq \ln \prod_{i=1}^n x_i^{t_i},$$

or equivalently

$$\sum_{i=1}^n t_i x_i \geq \prod_{i=1}^n x_i^{t_i},$$

for $x_1, \ldots, x_n \in (0, \infty)$.

In the special case where $t_1 = \cdots = t_n = \frac{1}{n}$, we have the inequality that relates the arithmetic to the geometric average on $n$ positive numbers:

$$\frac{x_1 + \cdots + x_n}{n} \geq \left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}}. \tag{B.4}$$

Let $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ be such that $\sum_{i=1}^n w_i = 1$. For $r \neq 0$, the ***w*-weighted mean of order** $r$ of a sequence of $n$ positive numbers $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n_{>0}$ is the number

$$\mu^r_{\mathbf{w}}(\mathbf{x}) = \left(\sum_{i=1}^n w_i x_i^r\right)^{\frac{1}{r}}.$$

Of course, $\mu^r_{\mathbf{w}}(\mathbf{x})$ is not defined for $r = 0$; we will give as special definition

$$\mu^0_{\mathbf{w}}(\mathbf{x}) = \lim_{r \to 0} \mu^r_{\mathbf{w}}(\mathbf{x}).$$

We have

$$\begin{aligned}
\lim_{r\to 0} \ln \mu^r_{\mathbf{w}}(\mathbf{x}) &= \lim_{r \to 0} \frac{\ln \sum_{i=1}^n w_i x_i^r}{r} \\
&= \lim_{r \to 0} \frac{\sum_{i=1}^n w_i x_i^r \ln x_i}{\sum_{i=1}^n w_i x_i^r} \\
&= \sum_{i=1}^n w_i \ln x_i \\
&= \ln \prod_{i=1}^n x_i^{w_i}.
\end{aligned}$$

Thus, if we define $\mu^0_{\mathbf{w}}(\mathbf{x}) = \prod_{i=1}^n x_i^{w_i}$, the weighted mean of order $r$ becomes a function continuous everywhere with respect to $r$.

For $w_1 = \cdots = w_n = \frac{1}{n}$, we have

$$\begin{aligned}\mu^{-1}_{\mathbf{w}}(\mathbf{x}) &= \frac{nx_1 \cdots x_n}{x_2 \cdots x_n + \cdots + x_1 \cdots x_{n-1}}\\ &\quad \text{(the harmonic average of } \mathbf{x}),\\ \mu^0_{\mathbf{w}}(\mathbf{x}) &= (x_1 \ldots x_n)^{\frac{1}{n}}\\ &\quad \text{(the geometric average of } \mathbf{x}),\\ \mu^1_{\mathbf{w}}(\mathbf{x}) &= \frac{x_1 + \cdots + x_n}{n}\\ &\quad \text{(the arithmetic average of } \mathbf{x}).\end{aligned}$$

**Theorem B.21.** *If $p < r$, we have $\mu^p_{\boldsymbol{w}}(\boldsymbol{x}) \le \mu^r_{\boldsymbol{w}}(\boldsymbol{x})$.*

*Proof.* There are three cases depending on the position of 0 relative to $p$ and $r$.

In the first case, suppose that $r > p > 0$. The function $f(x) = x^{\frac{r}{p}}$ is convex, so by Jensen's inequality applied to $x_1^p, \ldots, x_n^p$, we have

$$\left(\sum_{i=1}^n w_i x_i^p\right)^{\frac{r}{p}} \le \sum_{i=1}^n w_i x_i^r,$$

which implies

$$\left(\sum_{i=1}^n w_i x_i^p\right)^{\frac{1}{p}} \le \left(\sum_{i=1}^n w_i x_i^r\right)^{\frac{1}{r}},$$

which is the inequality of the theorem.

If $r > 0 > p$, the function $f(x) = x^{\frac{r}{p}}$ is again convex because $f''(x) = \frac{r}{p}\left(\frac{r}{p} - 1\right) x^{\frac{r}{p}-2} \ge 0$. Thus, the same argument works as in the previous case.

Finally, suppose that $0 > r > p$. Since $0 < \frac{r}{p} < 1$, the function $f(x) = x^{\frac{r}{p}}$ is concave. Thus, by Jensen's inequality,

$$\left(\sum_{i=1}^n w_i x_i^p\right)^{\frac{r}{p}} \ge \sum_{i=1}^n w_i x_i^r.$$

Since $\frac{1}{r} < 0$, we obtain again

$$\left(\sum_{i=1}^n w_i x_i^p\right)^{\frac{1}{p}} \le \left(\sum_{i=1}^n w_i x_i^r\right)^{\frac{1}{r}}.$$

□

# C

# Useful Integrals and Formulas

## C.1 Euler's Integrals

The integrals

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx,$$
$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx,$$

are known as *Euler's integral of the first type* and *Euler's integral of the second type*, respectively. We assume here that $a$ and $b$ are positive numbers to ensure that the integrals are convergent.

Replacing $x$ by $1-x$ yields the equality

$$B(a,b) = -\int_1^0 (1-x)^{a-1}(x)^{b-1}dx = B(b,a),$$

which shows that $B$ is symmetric.

Integrating $B(a,b)$ by parts, we obtain

$$\begin{aligned}
B(a,b) &= \int_0^1 x^{a-1}(1-x)^{b-1}dx \\
&= \int_0^1 (1-x)^{b-1}d\frac{x^a}{a} \\
&= \frac{x^a(1-x)^{1-b)}}{a}\Bigg|_0^1 + \frac{b-1}{a}\int_0^1 x^a(1-x)^{b-2}dx \\
&= \frac{b-1}{a}\int_0^1 x^{a-1}(1-x)^{b-2}dx - \frac{b-1}{a}\int_0^1 x^{a-1}(1-x)^{b-1}dx \\
&= \frac{b-1}{a}B(a,b-1) - \frac{b-1}{a}B(a,b),
\end{aligned}$$

which yields

$$B(a,b) = \frac{b-1}{a+b-1} B(a,b-1). \tag{C.1}$$

The symmetry of the function $B$ allows us to infer the formula

$$B(a,b) = \frac{a-1}{a+b-1} \cdot B(a-1,b).$$

If $b$ is a natural number $n$, a repeated application of Equality (C.1) allows us to write

$$B(a,n) = \frac{n-1}{a+n-1} \cdot \frac{n-2}{a+n-2} \cdots \frac{1}{a+1} \cdot B(a,1).$$

The last factor of this equality, $B(a,1)$, is easily seen to equal $\frac{1}{a}$. Thus,

$$B(a,n) = B(n,a) = \frac{1 \cdot 2 \cdot \cdots \cdot (n-1)}{a \cdot (a+1) \cdot \cdots \cdot (a+n-1)}.$$

If $a$ is also a natural number, $a = m \in \mathbb{N}$, then

$$B(m,n) = \frac{(n-1)!(m-1)!}{(m+n-1)!}.$$

Next, we show the connection between Euler's integral functions:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{C.2}$$

Changing the variable $x$ in the integral

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

by taking $x = ry$ with $r > 0$ gives

$$\Gamma(a) = r^a \int_0^\infty y^{a-1} e^{-ry} dy.$$

Replacing $a$ by $a+b$ and $r$ by $r+1$ yields the equality

$$\Gamma(a+b)(r+1)^{-(a+b)} = \int_0^\infty y^{a+b-1} e^{-(r+1)y} dy.$$

By multiplying both sides by $r^{a-1}$ and integrating, we have

$$\Gamma(a+b) \int_0^\infty r^{a-1}(r+1)^{-(a+b)} dr = \int_0^\infty r^{a-1} \left( \int_0^\infty y^{a+b-1} e^{-(r+1)y} dy \right) dr.$$

By the definition of $B$, the last equality can be written

$$\Gamma(a+b)B(a,b) = \int_0^\infty r^{a-1}\left(\int_0^\infty y^{a+b-1}e^{-(r+1)y}dy\right)dr.$$

By permuting the integrals from the right member (we omit the justification of this manipulation), the last equality can be written as

$$\Gamma(a+b)B(a,b) = \int_0^\infty y^{a+b-1}e^{-y}\left(\int_0^\infty r^{a-1}e^{-ry}dr\right)dy.$$

Note that $\int_0^\infty r^{a-1}e^{-ry}dr = \frac{\Gamma(a)}{y^a}$. Therefore,

$$\Gamma(a+b)B(a,b) = \int_0^\infty y^{a+b-1}e^{-y}\frac{\Gamma(a)}{y^a}dy = \int_0^\infty y^{b-1}e^{-y}\Gamma(a)dy = \Gamma(a)\Gamma(b),$$

which is Formula (C.2).

The $\Gamma$ function is a generalization of the factorial. Starting from the definition of $\Gamma$ and integrating by parts, we obtain

$$\Gamma(x) = \int_0^\infty x^{a-1}e^{-x}dx = \frac{x^a}{a}e^{-x}\Bigg|_0^\infty + \frac{1}{a}\int_0^\infty x^a e^{-x}dx = \frac{1}{a}\Gamma(a+1).$$

Thus, $\Gamma(a+1) = a\Gamma(a)$. Since $\Gamma(1) = \int_0^\infty e^{-x}dx = 1$, it is easy to see that $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$.

Using an argument from classical analysis it is possible to show that $\Gamma$ has derivatives of arbitrary order and that we can compute these derivatives by deriving the function under the integral sign. Namely, we can write:

$$\Gamma'(a) = \int_0^\infty x^{a-1}\ln x e^{-x}dx,$$

and, in general, $\Gamma^{(n)}(a) = \int_0^\infty x^{a-1}(\ln x)^n e^{-x}dx$. Thus, $\Gamma^{(2)}(a) > 0$, which shows that the first derivative is increasing.

Since $\Gamma(1) = \Gamma(2) = 1$, there exists $a \in [1,2]$ such that $\Gamma'(a) = 0$. For $0 < x < a$, we have $\Gamma'(x) \leq 0$, so $\Gamma$ is decreasing. For $x > a$, $\Gamma'(x) \geq 0$, so $\Gamma$ is increasing. It is easy to see that

$$\lim_{x\to 0+}\Gamma(x) = \frac{\Gamma(x+1)}{x} = \infty,$$

and $\lim_{x\to\infty}\Gamma(x) = \infty$.

An integral that is useful for a variety of applications is

$$I = \int_{\mathbb{R}} e^{-\frac{1}{2}t^2}dt.$$

We prove that $I = \sqrt{2\pi}$.

We can write

$$\begin{aligned} I^2 &= \int_{\mathbb{R}} e^{-\frac{1}{2}x^2} dx \cdot \int_{\mathbb{R}} e^{-\frac{1}{2}y^2} dy \\ &= \int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}} dxdy. \end{aligned}$$

Changing to polar coordinates by using the transformation

$$\begin{aligned} x &= \rho\cos\theta \\ y &= \rho\sin\theta, \end{aligned}$$

whose Jacobian is

$$\begin{vmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -\rho\sin\theta \\ \sin\theta & \rho\cos\theta \end{vmatrix} = \rho,$$

we have

$$\begin{aligned} I^2 &= \int_{\mathbb{R}^2} e^{-\frac{\rho^2}{2}} \rho d\rho d\theta \\ &= \int_0^{2\pi} d\theta \int_0^{\infty} e^{-\frac{\rho^2}{2}} \rho d\rho = 2\pi. \end{aligned}$$

Thus, $I = \sqrt{2\pi}$. Since $e^{-\frac{1}{2}t^2}$ is an even function, it follows that

$$\int_0^{\infty} e^{-\frac{1}{2}t^2} dt = \sqrt{\frac{\pi}{2}}.$$

Using this integral, we can compute the value of $\Gamma\left(\frac{1}{2}\right)$. Note that

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} \frac{e^{-x}}{\sqrt{x}} dx.$$

Applying the change of variable $x = \frac{t^2}{2}$, we have

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{2} \cdot \int_0^{\infty} e^{-\frac{1}{2}t^2} dt = \sqrt{\pi}. \tag{C.3}$$

The last equality allows us to compute the values of the form $\Gamma\left(\frac{2p+1}{2}\right)$. It is easy to see that

$$\Gamma\left(\frac{2p+1}{2}\right) = \frac{(2p-1)\cdot(2p-3)\cdots 3\cdot 1}{2^p}\sqrt{\pi} = \frac{(2p)!}{p!2^{2p}}\sqrt{\pi}. \tag{C.4}$$

## C.2 Wallis's Formula

The double factorial $n!!$ is defined by

$$n!! = \begin{cases} n(n-2)\cdots 4\cdot 2 & \text{if } n \text{ is even} \\ n(n-2)\cdots 3\cdot 1 & \text{if } n \text{ is odd.} \end{cases}$$

For example, $5!! = 5\cdot 3\cdot 1 = 15$ and $6!! = 6\cdot 4\cdot 2 = 48$.

Let $n \in \mathbb{N}$ and let $S_n = \int_0^{\frac{\pi}{2}} \sin^n x dx$ and $C_n = \int_0^{\frac{\pi}{2}} \cos^n x dx$. Integrating by parts, we obtain

$$\begin{aligned} S_n = \int_0^{\frac{\pi}{2}} \sin^{n-1} x d(-\cos x) &= \Big|_0^{\frac{\pi}{2}} + (m-1)\int_0^{\frac{\pi}{2}} \sin^{m-2} x \cos^2 x dx \\ &= (m-1)\int_0^{\frac{\pi}{2}} \sin^{m-2} x(1-\sin^2 x)dx \\ &= (m-1)S_{n-2} - (m-1)S_n, \end{aligned}$$

which implies

$$S_n = \frac{n-1}{n} S_{n-2}.$$

Note that $S_0 = \frac{\pi}{2}$ and that $S_1 = 1$. If $n$ is an even number, $n = 2p$, it follows that

$$S_{2p} = \frac{(2p-1)\cdot(2p-3)\cdots 3\cdot 1}{2p\cdot(2p-2)\cdots 4\cdot 2}\cdot\frac{\pi}{2} = \frac{(2p-1)!!}{2p!!}\frac{\pi}{2}.$$

If $n$ is an odd number, $n = 2p+1$, then

$$S_{2p+1} = \frac{2p\cdot(2p-2)\cdots 4\cdot 2}{(2p+1)\cdot(2p-1)\cdots 2\cdot 1} = \frac{(2p)!!}{(2p+1)!!}.$$

We can also show that $C_n = S_n$ for every $n \in \mathbb{N}$.

Using these results, we shall prove that

$$\lim_{n\to\infty}\left(\frac{2n!!}{(2n-1)!!}\right)^2 \cdot \frac{1}{2n+1} = \frac{\pi}{2}.$$

This equality is known as the *Wallis formula* and will help us prove Stirling's formula.

If $x \in (0, \frac{\pi}{2})$, we have $\sin^{2n+1} x < \sin^{2n} x < \sin^{2n-1} x$, which implies

$$\int_0^{\frac{pi}{2}} \sin^{2n+1} x dx < \int_0^{\frac{pi}{2}} \sin^{2n} x dx < \int_0^{\frac{pi}{2}} \sin^{2n-1} x dx.$$

Thus,

$$\frac{(2n)!!}{(2n+1)!!} < \frac{(2n-1)!!}{(2n)!!}\cdot\frac{\pi}{2} < \frac{(2n-2)!!}{(2n-1)!!}$$

or, equivalently,

$$\left(\frac{(2n)!!}{(2n-1)!!}\right)^2 \frac{1}{2n+1} < \frac{\pi}{2} < \left(\frac{(2n)!!}{(2n-1)!!}\right)^2 \frac{1}{2n}.$$

Note that

$$\begin{aligned}&\left(\frac{(2n)!!}{(2n-1)!!}\right)^2 \frac{1}{2n} - \left(\frac{(2n)!!}{(2n-1)!!}\right)^2 \frac{1}{2n+1}\\ &= \left(\frac{(2n)!!}{(2n-1)!!}\right)^2 \frac{1}{2n(2n+1)} < \frac{\pi}{4n},\end{aligned}$$

which gives Wallis' formula

$$\lim_{n\to\infty} \frac{1}{2n+1} \cdot \left(\frac{(2n)!!}{(2n-1)!!}\right)^2 = \frac{\pi}{2}.$$

Wallis' formula is equivalent to

$$\frac{\pi}{2} = \lim_{n\to\infty} \frac{1}{2n+1} \cdot \left(\frac{2^{2n}(n!)^2}{(2n)!}\right)^2. \tag{C.5}$$

## C.3 Stirling's Formula

The starting point for proving Stirling's formula is the power series

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1}\frac{x^n}{n} + \cdots,$$

which is convergent for $x \in [-1, 1]$.

Replacing $x$ by $-x$, we obtain

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots - \frac{x^n}{n} + \cdots,$$

which allows us to write

$$\ln\frac{1+x}{1-x} = 2x\left(1 + \frac{x^2}{3} + \frac{x^4}{5} + \cdots + \frac{x^{2n}}{2n+1} + \cdots\right).$$

Choosing $x = \frac{1}{2n+1}$, we have

$$\ln\frac{n+1}{n} = \frac{2}{2n+1}\left(1 + \frac{1}{3}\cdot\frac{1}{(2n+1)^2} + \frac{1}{5}\cdot\frac{1}{(2n+1)^4} + \cdots\right)$$

or, equivalently,

$$\left(n+\frac{1}{2}\right)\ln\left(1+\frac{1}{n}\right)=1+\frac{1}{3}\cdot\frac{1}{(2n+1)^2}+\frac{1}{5}\frac{1}{(2n+1)^4}+\cdots.$$

Note that

$$\begin{aligned}&1+\frac{1}{3}\cdot\frac{1}{(2n+1)^2}+\frac{1}{5}\cdot\frac{1}{(2n+1)^4}+\cdots\\&\le 1+\frac{1}{3}\left(\frac{1}{(2n+1)^2}+\frac{1}{(2n+1)^4}+\cdots\right)\\&=1+\frac{1}{12n(n+1)},\end{aligned}$$

which implies that

$$1<\left(n+\frac{1}{2}\right)\ln\left(1+\frac{1}{n}\right)<1+\frac{1}{12n(n+1)}.$$

The last inequalities can be written as

$$e<\left(1+\frac{1}{n}\right)^{n+\frac{1}{2}}<e^{1+\frac{1}{12n(n+1)}}=e\cdot\frac{e^{\frac{1}{12n}}}{e^{\frac{1}{12(n+1)}}}.$$

Define the sequence

$$x_n=\frac{n!e^n}{n^{n+\frac{1}{2}}}$$

for $n\in\mathbb{N}$. It is easy to see that

$$\frac{x_n}{x_{n+1}}=\frac{1}{e}\left(1+\frac{1}{n}\right)^{n+\frac{1}{2}}.$$

Thus, we have

$$1<\frac{x_n}{x_{n+1}}<\frac{e^{\frac{1}{12n}}}{e^{\frac{1}{12(n+1)}}}.$$

This double inequality implies that $(x_n)$ is a decreasing sequence and that the sequence defined by $z_n=x_ne^{-\frac{1}{12n}}$ is an increasing sequence. Since $(x_n)$ has 0 as a lower bound, it follows that $\ell=\lim_{n\to\infty}x_n$ exists. This also implies that $\ell=\lim_{n\to\infty}z_n$ and, since $z_n$ is increasing, it also follows that $z_n<\ell<x_n$. The inequality

$$x_ne^{-\frac{1}{12n}}<\ell<x_n$$

implies that there exists $\theta_n\in(0,1)$ such that $\ell=x_ne^{-\frac{\theta_n}{12n}}$, or $x_n=\ell e^{\frac{\theta_n}{12n}}$. By the definition of $x_n$, we have

$$\frac{n!e^n}{n^{n+\frac{1}{2}}}=\ell e^{\frac{\theta_n}{12n}},$$

which yields the equality

$$n! = \ell\sqrt{n}\left(\frac{n}{e}\right)^n e^{\frac{\theta_n}{12n}} \tag{C.6}$$

To determine $\ell$ we use Equality (C.5) and replace the factorials in this equality using Equality (C.6). Observe that

$$\frac{2^{2n}(n!)^2}{(2n)!} = \ell\sqrt{\frac{n}{2}}e^{\frac{4\theta_n-\theta_{2n}}{24n}},$$

which allows us to write

$$\frac{\pi}{2} = \lim_{n\to\infty}\frac{\ell^2}{2n+1}\cdot\frac{n}{2}e^{\frac{4\theta_n-\theta_{2n}}{12n}} = \frac{\ell^2}{4}.$$

Thus, $\ell = \sqrt{2\pi}$, and this yields Stirling's formula

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{\theta_n}{12n}}. \tag{C.7}$$

## C.4 The Volume of an *n*-Dimensional Sphere

A closed sphere centered in $(0,\ldots,0)$ and having the radius $R$ in $\mathbb{R}^n$ is defined as the set of points:

$$S_n(R) = \left\{(x_1,\ldots,x_n)\in\mathbb{R}^n \mid \sum_{i=1}^n x_i^2 = 1\right\}.$$

The volume of this sphere is denoted by $V_n(R)$.

We approximate the volume of an $n$-dimensional sphere of radius $R$ as a sequence of $n-1$-dimensional spheres of radius $r(u) = \sqrt{R^2-u_2}$, where $u$ varies between $-R$ and $R$. This allows us to write

$$V_{n+1}(R) = \int_{-R}^{R} V_n(r(u))du.$$

We seek $V_n(R)$ as a number of the form $V_n(R) = k_n R^n$. Thus, we have

$$\begin{aligned}
V_{n+1}(R) &= k_n\int_{-R}^{R}(r(u))^n du\\
&= k_n\int_{-R}^{R}(R^2-u^2)^{\frac{n}{2}}du\\
&= k_n R^n\int_{-R}^{R}\left(1-\left(\frac{u}{R}\right)^2\right)^{\frac{n}{2}}du\\
&= V_n(R)\int_{-R}^{R}\left(1-\left(\frac{u}{R}\right)^2\right)^{\frac{n}{2}}du\\
&= RV_n(R)\int_{-1}^{1}(1-x^2)^{\frac{n}{2}}dx.
\end{aligned}$$

In turn, this yields the recurrence

$$k_{n+1} = k_n \int_{-1}^{1} (1 - x^2)^{\frac{n}{2}} dx.$$

Note that

$$\int_{-1}^{1} (1 - x^2)^{\frac{n}{2}} dx = 2 \cdot \int_{0}^{1} (1 - x^2)^{\frac{n}{2}} dx$$

because the function $(1 - x^2)^{\frac{n}{2}}$ is even. To compute the latest integral, substitute $u = x^2$. We obtain

$$\int_{0}^{1} (1 - x^2)^{\frac{n}{2}} dx = \frac{1}{2} \int_{0}^{1} u^{-\frac{1}{2}} (1 - u)^{\frac{n}{2}} du,$$

which equals $\frac{1}{2} \cdot B(\frac{1}{2}, \frac{n}{2} + 1)$. Using the $\Gamma$ function, the integral can be written as

$$\int_{0}^{1} (1 - x^2)^{\frac{n}{2}} dx = \frac{1}{2} \cdot \frac{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2} + 1)}{\Gamma\left(\frac{n}{2} + \frac{3}{2}\right)}.$$

Thus,

$$k_{n+1} = k_n \frac{\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n}{2} + 1\right)}{\Gamma\left(\frac{n+1}{2} + 1\right)}.$$

Since $k_1 = 2$, this implies

$$\begin{aligned} k_n &= 2 \left(\Gamma\left(\frac{1}{2}\right)\right)^{n-1} \frac{\Gamma\left(\frac{1}{2} + 1\right)}{\Gamma\left(\frac{n}{2} + 1\right)} \\ &= \left(\Gamma\left(\frac{1}{2}\right)\right)^{n} \frac{1}{\Gamma\left(\frac{n}{2} + 1\right)} \\ &= \pi^{\frac{n}{2}} \frac{1}{\Gamma\left(\frac{n}{2} + 1\right)}. \end{aligned}$$

Thus, the volume of the $n$-dimensional sphere of radius $R$ equals

$$\frac{\pi^{\frac{n}{2}} R^n}{\Gamma\left(\frac{n}{2} + 1\right)}.$$

For $n = 1, 2, 3$, by applying Formula (C.4), we obtain the well-known values $2R$,$\pi R^2$, and $\frac{4\pi R^3}{3}$, respectively. For $n = 4$, the volume of the sphere is $\frac{\pi^2 R^4}{2}$.

# D

# A Characterization of a Function

The goal of this section is to show that if $h : \mathbb{N} \longrightarrow \mathbb{R}$ is an increasing function such that $h(2) = 2$ and $h(mn) = mh(n) + nh(m)$ for $m, n \in \mathbb{N}$, then $h(n) = n \log_2 n$ for every $n \in \mathbb{N}_1$.

For a number $x \in \mathbb{R}$, we denote the largest integer that is less than or equal to $x$ by $\lfloor x \rfloor$; the *fractional part* of $x$ will be denoted by $\langle x \rangle$, where $\langle x \rangle = x - \lfloor x \rfloor$.

The following technical lemma is a special case of a result of Dirichlet (see[151], p.235).

**Lemma D.1.** *Let $\alpha$ be a real number and let $q$ be a positive integer. There exists $m$ such that $1 \leq m < q$ and an integer $n$ such that $|m\alpha - n| < \dfrac{1}{q} < \dfrac{1}{m}$.*

*Proof.* Divide the set $I = \{x \in \mathbb{R} | 0 \leq x < 1\}$ into $q$ equal subintervals $\left[\frac{i-1}{q}, \frac{i}{q}\right)$ for $1 \leq i \leq q$. Of the $q + 1$ numbers $\langle n\alpha \rangle$, where $0 \leq n \leq q$, at least two, say $\langle n_1\alpha \rangle$ and $\langle n_2\alpha \rangle$ are in the same subinterval. This means $|\langle n_1\alpha \rangle - \langle n_2\alpha \rangle| < \frac{1}{q}$. Setting $\lfloor n_1\alpha \rfloor = n'$ and $\lfloor n_2\alpha \rfloor = n''$ we obtain $|n_1\alpha - \lfloor n_1\alpha \rfloor - n_2\alpha + \lfloor n_2\alpha \rfloor| < \frac{1}{q}$, or $|(n_1 - n_2)\alpha - (n' - n'')| < \frac{1}{q}$. We can take $m = n_1 - n_2 \geq 1$ and $n = n' - m''$ in order to obtain the desired inequality. □

**Lemma D.2.** *If $n \leq m\epsilon < n + \epsilon$, then there exist $m'$ and $n'$ such that $n' - \epsilon < m'\epsilon < n'$.*

*Similarly, if $n - \epsilon < m\epsilon < n$, there exist $n''$ and $m''$ such that $n'' \leq m''\epsilon < n'' + \epsilon$.*

*Proof.* Let $\theta, \sigma > 0$. We have $\theta n + \sigma\alpha \leq (\theta m + \sigma)\alpha < \theta n + \theta\epsilon + \sigma\alpha$. Choose $\theta > \max\{1, 1/\epsilon\}$. Under this choice, the interval $[\theta n + \sigma\alpha, \theta n + \theta\epsilon + \sigma\alpha)$ is of length greater than 1, and therefore there is an integer $m'$ in this interval for any choice of $\sigma$. This allows us to choose $\theta$ and $\sigma$ such that $\theta m + \sigma = m'$

and $\theta n + \theta\epsilon + \sigma\alpha = n'$. Note that if we make these choices, then $\theta n + \sigma\alpha = n' - \theta\epsilon > n' - \epsilon$ and therefore $n' - \epsilon \leq m\alpha < n'$.

The second part of the argument is similar and is left to the reader. □

**Lemma D.3.** *If $h : \mathbb{N} \longrightarrow \mathbb{R}$ is a function such that*

$$h(mn) = mh(n) + nh(m)$$

*for every $m, n \in \mathbb{N}$, then $h(p^k) = kp^{k-1}h(p)$ for every $p, k \in \mathbb{N}$ and $k \geq 1$.*

*Proof.* The argument is by induction on $k$ and is left to the reader. □

**Lemma D.4.** *If $h : \mathbb{N} \longrightarrow \mathbb{R}$ is a function such that*

$$h(mn) = mh(n) + nh(m)$$

*for every $m, n \in \mathbf{N}$, then*

$$h(p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}) = p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n} \sum_{1 \leq i \leq n} \frac{k_i h(p_i)}{p_i}.$$

Let $\ell : \mathbb{N} \longrightarrow \mathbb{R}$ be the function given by

$$\ell(n) = \begin{cases} 0 & \text{if } n = 0, \\ \frac{h(n)}{n} & \text{if } n > 0. \end{cases}$$

Note that $\ell(mn) = \ell(m) + \ell(n)$ and therefore

$$\ell(p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}) = \ell(p_1^{k_1}) + \ell(p_2^{k_2}) + \cdots + \ell(p_n^{k_n}).$$

Since $\ell(p^k) = \dfrac{k}{p} h(p)$ because of Lemma D.3, we obtain

$$\ell(p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}) = \sum_{1 \leq i \leq n} \frac{k_i h(p_i)}{p_i},$$

which immediately gives the equality of the lemma. □

**Theorem D.5.** *Let $h : \mathbb{N} \longrightarrow \mathbb{R}$ be a function such that $h(p) = p \log p$ if $p = 1$ or if $p$ is prime. If $h(mn) = mh(n) + nh(m)$ for every $m, n \in \mathbb{N}$, then $h(n) = n \log n$ for every $n \in \mathbb{N}$, $n \geq 1$.*

*Proof.* Since every positive integer $n$ other than 1 can be written uniquely as a product of powers of primes $n = p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}$, we have

$$\begin{aligned} h(n) &= p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n} \sum_{1 \leq i \leq n} \frac{k_i h(p_i)}{p_i} \\ &= n \sum_{1 \leq i \leq n} k_i \log p_i \\ &= n \log n \end{aligned}$$

for $n \geq 2$. □

**Theorem D.6.** *Let $h : \mathbb{N} \longrightarrow \mathbb{R}$ be an increasing function such that $h(mn) = mh(n) + nh(m)$ for every $m, n \in \mathbb{N}$. If $h(2) = 2$, then $h(n) = n \log_2 n$ for $n \in \mathbb{N}$.*

*Proof.* Define the function $b : \{n \in \mathbb{N} | n > 1\} \longrightarrow \mathbb{R}$ by $b(n) = h(n)/(n * \log n)$.

We shall prove initially that if $p > 2$ is a prime number, then $b(p) \geq 1$. Let $\epsilon > 0$ be a real number. Taking $q < \frac{1}{\epsilon}$ in Lemma D.1, we obtain the existence of $m, n \in \mathbb{N}$ such that $|m\alpha - n| < \epsilon$. In other words, we have $n - \epsilon < m\alpha < n + \epsilon$. If $n < m\alpha < n + \epsilon$, then by Lemma D.2 there are $m', n'$ such that $n' - \epsilon < m'\epsilon < n'$. If $n - \epsilon < m\alpha < n$, then the same lemma implies the existence of $n'', m''$ such that $n'' \leq m''\epsilon < n'' + \epsilon$.

If we choose $\alpha = \log p$, then we may assume that there are $m, n \in \mathbb{N}$, $m, n \geq 1$ such that $n \leq m \log p < n + \epsilon$. Equivalently, we have $2^n \leq p^m < 2^n\ 2^\epsilon$. Since $h$ is an increasing function, we obtain $n2^n \leq h(p^m)$, or $n2^n \leq mp^{m-1}h(p)$. Because of the definition of $b$ we have $n2^n \leq mp^m b(p) \log p$, or $n2^n \leq b(p)p^m \log p^m$. In view of the previous inequality, this implies

$$n2^n \leq b(p)2^n 2^\epsilon (n + \epsilon)$$

or, equivalently,

$$b(p) \geq \frac{n}{2^\epsilon (n + \epsilon)}.$$

Taking $\epsilon \to 0$, we obtain $b(p) \geq 1$.

Similarly, there exists a number $m \in \mathbb{N}$ such that $n - \epsilon < m \log p \leq n$. A similar argument that makes use of Lemma D.2 shows that $b(p) \leq 1$, so $b(p) = 1$, which proves that $h(p) = p \log p$ for every prime $p$. □

# References


1. J.-M. Adamo. *Data Mining for Association Rules and Sequential Patterns*. Springer-Verlag, New York, 2001.
2. Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. Depth first generation of long patterns. In R. Bayardo, R. Ramakrishnan, and S. J. Stolfo, editors, *Proceedings of the 6th Conference on Knowledge Discovery in Data, Boston, MA*, pages 108–118. ACM, New York, 2000.
3. Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. *Journal of Parallel and Distributed Computing*, 61(3):350–371, 2001.
4. C. C. Aggarwal and P. S. Yu. Mining associations with the collective strength approach. *IEEE Transactions on Knowledge and Data Engineering*, 13:863–873, 2001.
5. Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 International Conference on Management of Data*, pages 207–216, Washington, D.C., 1993. ACM, New York.
6. P. Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33:233–282, 1983.
7. R. Bayardo and R. Agrawal. Mining the most interesting rules. In S. Chaudhuri and D. Madigan, editors, *Proceedings of the 5th KDD, San Diego, CA*, pages 145–153. ACM, New York, 1999.
8. R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ, 1961.
9. C. Berge. *Hypergraphes*. North-Holland, Amsterdam, 1989.
10. C. Berge. *The Theory of Graphs*. Dover, Mineola, NY, 2001.
11. P. Berkhin. A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping Multidimensional Data – Recent Advances in Clustering*, pages 25–72. Springer-Verlag, Berlin, 2006.
12. P. Berkhin and J. Becher. Learning simple relations: Theory and applications. In R. L. Grossman, J. Han, V. Kumar, H. Mannila, and R. Motwani, editors, *Proceedings of the Second SIAM International Conference on Data Mining, Arlington, Virginia*, pages 420–436. SIAM, 2002.
13. P. Billingsley. *Ergodic Theory and Information*. John Wiley, New York, 1965.

14. G. Birkhoff. *Lattice Theory*. American Mathematical Society, Providence, RI, third edition, 1973.
15. A. Björklund and T. Husfeldt. Inclusion-exclusion algorithms for counting set partitions. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, Berkeley, CA*, pages 575–582. IEEE Computer Society Press, Los Alamitos, CA, 2006.
16. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.
17. C. Böhm, S. Berthold, and D. A. Keim. Searching in high-dimensional spaces – index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33:322–373, 2001.
18. B. Bollobás. *Graph Theory – An Introductory Course*. Springer-Verlag, New York, 1979.
19. Z. Bonikowski. A certain conception of the calculus of rough sets. *Notre Dame Journal of Formal Logic*, 33:412–421, 1992.
20. E. Boros, P. L. Hammer, T. Ibaraki, and A. Kogan. A logical analysis of numerical data. *Mathematical Programming*, 79:163–190, 1997.
21. E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik. An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering*, 12:292–306, 2000.
22. U. Brandes, D. Delling, M. Gaertler, R.Görke, M. Hoefer, Z. Nikolski, and D. Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, 2008.
23. S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In J. Pekham, editor, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 265–276, Tucson, AZ, 1997. ACM, New York.
24. P. Buneman. A note on metric properties of trees. *Journal of Combinatorial Theory (B)*, 17:48–50, 1974.
25. W. A. Burkhard and R. M. Keller. Some approaches to best-match file searching. *Communications of the ACM*, 16:230–236, 1973.
26. T. Calders. *Axiomatization and Deduction Rules for the Frequency of Item Sets*. PhD thesis, Universiteit Antwerpen, 2003.
27. E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33:273–321, 2001.
28. K. L. Clarkson. Nearest-neighbor searching and metric space dimensions. In G. Shakhnarovich, T. Darrell, and P. Indyk, editors, *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, pages 15–59. MIT Press, Cambridge, MA, 2006.
29. Kenneth Clarkson. Nearest neighbor queries in metric spaces. In *Proceedings of the 29th ACM Symposium on Theory of Computation*, pages 609–617, El Paso, TX, 1997. ACM, NY.
30. E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13:377–387, 1970.
31. E. F. Codd. *The Relational Model for Database Management, Version 2*. Addison-Wesley, Reading, MA, 1990.
32. P. M. Cohn. *Universal Algebra*. D. Reidel, Dordrecht, 1981.
33. M. M. Dalkilic and E. L. Robertson. Information dependencies. Technical Report TR531, Indiana University, 1999.

34. Z. Daróczy. Generalized information functions. *Information and Control*, 16:36–51, 1970.
35. C. J. Date. *An Introduction to Database Systems*. Addison-Wesley, Boston, eighth edition, 2003.
36. E. Deza and M. M. Deza. *Dictionary of Distances*. Elsevier, Amsterdam, 2006.
37. M. M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer-Verlag, Berlin, 1997.
38. R. Diestel. *Graph Theory*. Springer-Verlag, Berlin, 2005.
39. G. A. Dirac. Some theorems on abstract graphs. *Proceedings of the London Mathematical Society*, 2:69–81, 1952.
40. J. Dixmier. *General Topology*. Springer-Verlag, New York, 1984.
41. R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge, 1999.
42. I. Düntsch and G. Gediga. *Rough Sets Analysis - A Road to Non-invasive Knowledge Discovery*. Methoδos, Bangor, UK, 2000.
43. B. Dushnik and E. W. Miller. Partially ordered sets. *American Journal of Mathematics*, 63:600–610, 1941.
44. G. Edgar. *Measure, Topology, and Fractal Geometry*. Springer-Verlag, New York, 1990.
45. G. Edgar. *Integral, Probability, and Fractal Measures*. Springer-Verlag, New York, 1998.
46. M. Eisenberg. *Topology*. Holt, Rinehart and Winston, Inc., New York, 1974.
47. R. Engelking and K. Siekluchi. *Topology - A Geometric Approach*. Heldermann Verlag, Berlin, 1992.
48. K. Falconer. *Fractal Geometry*. Wiley, New York, second edition, 2003.
49. C. Faloutsos and K. I. Lin. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In Michael J. Carey and Donovan A. Schneider, editors, *Proceedings of SIGMOD, San Jose, California*, pages 163–174. ACM Press, NY, 1995.
50. A. Faragó, T. Linder, and G. Lugosi. Fast nearest-neighbor search in dissimilarity spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:957–962, 1993.
51. P. A. Fejer and D. A. Simovici. *Mathematical Foundations of Computer Science*, volume 1. Springer-Verlag, New York, 1991.
52. J. B. Fraleigh. *A First Course in Abstract Algebra*. Addison-Wesley, Reading, MA, 1982.
53. M. Fréchet. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo*, 22:1–47, 1906.
54. M. Fréchet. Les dimensions d'un ensemble abstrait. *Mathematische Annalen*, 68:145–178, 1910.
55. F. R. Gantmacher. *The Theory of Matrices (2 vols.)*. American Mathematical Society, Providence, RI, 1977.
56. J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18:54–64, 1969.
57. R. L. Graham, D. E. Knuth, and O. Patashnick. *Concrete Mathematics - A Foundation for Computer Science*. Addison-Wesley, Reading, MA, 1989.
58. W. Greub. *Linear Algebra*. Springer-Verlag, New York, fourth edition, 1981.
59. B. Grünbaum. *Convex Polytopes*. Wiley Interscience, London, 1967.
60. S. L. Hakimi. On the realizability of a set of integers as degrees of the vertices of a graph. *SIAM Journal on Applied Mathematics*, 10:496–506, 1962.

61. P. R. Halmos. *Measure Theory*. Van Nostrand, New York, 1959.
62. P. R. Halmos. *Naive Set Theory*. Springer-Verlag, New York, 1974.
63. P. L. Hammer and S. Rudeanu. *Méthodes Booléennes en Recherche Operationnelle*. Dunod, Paris, 1970.
64. J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, *Proceedings of the ACM-SIGMOD International Conference on Management of Data, Dallas, TX*, pages 1–12. ACM, New York, 2000.
65. F. Harary. *Graph Theory*. Addison-Wesley, Reading, MA, 1971.
66. V. Havel. A remark on the existence of finite graphs. *Časopis pro Pěstování Matematiky*, 80:477–480, 1955.
67. J. H. Havrda and F. Charvat. Quantification methods of classification processes: Concepts of structural $\alpha$-entropy. *Kybernetica*, 3:30–35, 1967.
68. R. Hilderman and H. Hamilton. Knowledge discovery and interestingness measures: A survey. Technical Report CS 99-04, Department of Computer Science, University of Regina, 1999.
69. C. M. Huang, Q. Bi, G. S. Stiles, and R. W. Harris. Fast full search equivalent encoding algorithms for immage compression using vector quantization. *IEEE Transactions on Image Processing*, 1:413–416, 1992.
70. Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen. Efficient discovery of functional and approximate dependencies using partitions (extended version). TR C-79, University of Helsinki, Department of Computer Science, Helsinki, Finland, 1997.
71. W. Hurewicz and H. Wallman. *Dimension Theory*. Princeton University Press, Princeton, 1948.
72. A. K. Jain and R. C. Dubes. *Algorithm for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, 1988.
73. A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31:264–323, 1999.
74. S. Jaroszewicz and D. Simovici. Interestingness of frequent item sets using bayesian networks as background knowledge. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA*, pages 178–186. ACM, New York, 2004.
75. S. Jaroszewicz and D. A. Simovici. On axiomatization of conditional entropy. In *Proceedings of the 29th International Symposium for Multiple-Valued Logic, Freiburg, Germany*, pages 24–31. IEEE Computer Society, Los Alamitos, CA, 1999.
76. L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data – An Introduction to Cluster Analysis*. Wiley Interscience, New York, 1990.
77. J. L. Kelley. *General Topology*. Van Nostrand, Princeton, NJ, 1955.
78. D. W. Khan. *Topology - An Introduction to Point-Set and Algebraic Areas*. Dover, New York, second edition, 1995.
79. J. Kivinen and H. Mannila. Approximate dependency inference from relations. *Theoretical Computer Science*, 149:129–149, 1995.
80. J. Kleinberg. An impossibility theorem for clustering. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15, Vancouver, Canada, 2002*, pages 446–453. MIT Press, Cambridge, MA, 2003.

81. Jan Komorowski, Zdzislaw Pawlak, Lech Polkowski, and Andrzej Skowron. Rough sets: A tutorial. In S. K. Pal and A. Skowron, editors, *Rough Sets Hybridization: A New Trend in Decision Making*, pages 3–98. Springer-Verlag Telos, 1999.
82. M. Köppen. The curse of dimensionality. 5th Online World Conference on Soft Computing in Industrial Applications (WSC5), 2000.
83. Flip Korn, Bernd-Uwe Pagel, and Christos Faloutsos. On the "dimensionality curse" and the "self similarity blessing". *IEEE Transactions on Knowledge and Data Engineering*, 13:96–111, 2001.
84. T. Kurita. An efficient agglomerative clustering algorithm using a heap. *Pattern Recognition*, 24:205–209, 1991.
85. P. D. Lax. *Linear Algebra and Its Applications*. Wiley-International, Hoboken, NJ, 2007.
86. M. Ledoux and M. Talagrand. *Isoperimetry and Processes in Probability in Banach Spaces*. Springer-Verlag, Berlin, 2002.
87. T. T. Lee. An information-theoretic analysis of relational databases. *IEEE Transactions on Software Engineering*, SE-13:1049–1061, 1997.
88. V. I. Levenshtein. Binary code capable of correcting deletions, insertions and substitutions. *Cybernetics and Control Theory*, 163:845–848, 1966.
89. E. H. Lieb and M. Ruskai. Proof of the strong subadditivity of quantum-mechanical entropy. *Journal of Mathematical Physics*, 14:1938–1941, 1973.
90. J. B. Listing. *Vorstudien zur Topologie*. Vanderhoek and Ruprecht, Göttingen, 1848.
91. D. Lubell. A short proof of Sperner Theorem. *Journal of Combinatorial Theory*, 1:299, 1966.
92. D. Maier. *The Theory of Relational Databases*. Computer Science Press, Rockville, MD, 1983.
93. F. M. Malvestuto. Statistical treatment of the information content of a database. *Information Systems*, 11:211–223, 1986.
94. B. Mandelbrot. *The Fractal Geometry of Nature*. W. H. Freeman, New York, 1982.
95. H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. Technical Report C-1997-8, University of Helsinki, 1997.
96. J. Marica and J. Schönheim. Differences of sets and a problem of Graham. *Canadian Mathematical Bulletin*, 12:635–637, 1969.
97. M. Meilă. Comparing clusterings: An axiomatic view. In L. DeRaedt and S. Wrobel, editors, *International Conference on Machine Learning, Bonn, Germany*, pages 577–584. ACM, New York, 2005.
98. L. D. Meshalkin. A generalization of Sperner's Theorem on the number of subsets of a finite set. *Theory of Probability and Its Applications (in Russian)*, 8:203–204, 1963.
99. T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
100. M. E. Munroe. *Introduction to Measure and Integration*. Addison-Wesley, Reading, MA, 1959.
101. M. T. Orchard. A fast nearest-neighbor search algorithm. In *International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 2297–3000, 1991.
102. O. Ore. Arc coverings of graphs. *Annali di Matematica Pura ed Applicata*, 55:315–322, 1961.

103. O. Ore. Hamilton connected graphs. *Journal de Mathématiques Pures et Appliquées*, 42:21–27, 1963.
104. Bernd-Uwe Pagel, Flip Korn, and Christos Faloutsos. Deflating the dimensionality curse using multiple fractal dimensions. In *International Conference on Data Engineering*, pages 589–598, 2000.
105. Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In C. Beeri and P. Buneman, editors, *Database Theory – ICDT'99, Jerusalem, Israel*, volume 1540 of *Lecture Notes in Computer Science*, pages 398–416. Springer-Verlag, Berlin, 1999.
106. Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishing, Dordrecht, 1991.
107. G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. MIT Press, Cambridge, MA, 1991.
108. D. Pollard. *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics, Hayward, CA, 1990.
109. H. Prüfer. Neuer Beweis eines Satzes über Permutationen. *Archiv für Mathematik und Physik*, 27:142–144, 1918.
110. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.
111. W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 61:846–850, 1971.
112. S. Ray and R. Turi. Determination of number of clusters in $k$-means clustering in colour image segmentation. In *Proceedings of the 4th Int. Conference on Advances in Pattern Recognition and Digital Technology*, pages 137–143, New Delhi, India, 1984. Narosa.
113. C. Reischer, D. A. Simovici, and M. Lambert. *Introduction aux Structures Algébriques*. Éditions du Renouveau Pédagogique, Montréal, 1992.
114. S. Rudeanu. *Boolean Functions and Equations*. North-Holland, Amsterdam, 1974.
115. S. Rudeanu. *Lattice Functions and Equations*. Springer-Verlag, London, 2001.
116. R. Rymon. Search through systematic set enumeration. In Bernhard Nebel, Charles Rich, and William R. Swartout, editors, *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning, Cambridge, MA*, pages 539–550. Morgan Kaufmann, San Mateo, CA, 1992.
117. N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13:145–147, 1972.
118. B. Sayrafi. *A Measure-Theoretic Framework for Constraints and Bounds on Measurements of Data*. PhD thesis, Indiana University, 2005.
119. B. Sayrafi and D. van Gucht. Differential constraints. In Chen Li, editor, *Principles of Database Systems, Baltimore, MD*, pages 348–357. ACM New York, 2005.
120. B. Sayrafi, D. van Gucht, and M. Gyssens. Measures in databases and datamining. Technical Report TR602, Indiana University, 2004.
121. G. Schay. *Introduction to Linear Algebra*. Jones and Bartlett, Sudbury, MA, 1997.
122. D. Shasha and T. L. Wang. New techniques for best-match retrieval. *ACM Transactions on Information Systems*, 8:140–158, 1990.

123. D. Simovici and S. Jaroszewicz. Generalized conditional entropy and decision trees. In *Extraction et Gestion des connaissances – EGC 2003*, pages 363–380. Lavoisier, Paris, 2003.
124. D. A. Simovici, D. Cristofor, and L. Cristofor. Impurity measures in databases. *Acta Informatica*, 38:307–324, 2002.
125. D. A. Simovici and S. Jaroszewicz. On information-theoretical aspects of relational databases. In C. Calude and G. Paun, editors, *Finite versus Infinite*, pages 301–321. Springer-Verlag, London, 2000.
126. D. A. Simovici and R. L. Tenney. *Relational Database Systems*. Academic Press, New York, 1995.
127. E. Sperner. Ein Satz über Untermengen einer endlichen Menge. *Mathematische Zeitschrift*, 27:544–548, 1928.
128. R. P. Stanley. *Enumerative Combinatorics*, volume 1. Cambridge University Press, Cambridge, 1997.
129. R. P. Stanley. *Enumerative Combinatorics*, volume 2. Cambridge University Press, Cambridge, 1999.
130. M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In M. Grobelnik, D. Mladenic, and N. Milic-Freyling, editors, *KDD Workshop on Text Mining, Boston, MA*, 2000.
131. P. Suppes. *Axiomatic Set Theory*. Dover, New York, 1972.
132. P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, Reading, MA, 2005.
133. R. E. Tarjan. *Data Structures and Network Algorithms*. SIAM, Philadelphia, 1983.
134. M. E. Taylor. *Measure Theory and Integration*. American Mathematical Society, Providence, R.I., 2006.
135. W. M. Trotter. *Combinatorics and Partially Ordered Sets*. The Johns Hopkins University Press, Baltimore, 1992.
136. W. T. Trotter. Partially ordered sets. In R. L. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics*, pages 433–480. The MIT Press, Cambridge, MA, 1995.
137. J. D. Ullman. *Database and Knowledge-Base Systems (2 vols.)*. Computer Science Press, Rockville, MD, 1988.
138. J. H. van Lint and R. M. Wilson. *A Course in Combinatorics*. Cambridge University Press, Cambridge, second edition, 2002.
139. V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Applications*, 16:264–280, 1971.
140. E. Vidal. An algorithm for finding nearest neighbours in (approximately average constant time. *Pattern Recognition Letters*, 4:145–157, 1986.
141. E. Vidal. New formulation and improvements of the nearest neighbours in approximating end eliminating search algorithms (AESA). *Pattern Recognition Letters*, 15:1–7, 1994.
142. M. Vidyasagar. *Learning and Generalization with Applications to Neural Networks*. Springer-Verlag, London, 2003.
143. T. L. Wang and D. Shasha. Query processing for distance metrics. In D. McLeod, R. Sacks-Davis, and H.-J.Schek, editors, *Proceedings of the 16th VLDB Conference, Brisbane, Australia*, pages 602–613. Morgan Kaufmann, San Francisco, CA, 1990.

144. R. S. Wenocur and R. M. Dudley. Some special Vapnik-Chervonenkis classes. *Discrete Mathematics*, 33:313–318, 1981.
145. K. Yamamoto. Logarithmic order of free distributive lattices. *Journal of Mathematical Society of Japan*, 6:343–353, 1954.
146. Y. Y. Yao. Two views of the theory of rough sets in finite universes. *International Journal of Approximate Reasoning*, 15:291–317, 1996.
147. C. T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, C-20:68–86, 1971.
148. M. J. Zaki and C.J. Hsiao. Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering*, 17:462–478, 2005.
149. K. Zakloutal, M. H. Johnson, and R. E. Ladner. Nearest neighbor search for data compression. In M. H. Goldwasser, D. S. Johnson, and C.C. McGeogh, editors, *DIMACS Series in Discrete Mathematics: Data Structures, Near Neighbor Searches, and Methodology: Fifth and Sixth Implementation Challenges*, pages 69–86. American Mathematical Society, Providence, RI, 2002.
150. Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.
151. A. Zygmund. *Trigonometric Series*. Cambridge University Press, Cambridge, second edition, 1990.

# Topic Index


Abelian groups 60
absorption laws 177
accumulation point 235, 431
accuracy of approximation 336
additive inverse of an element 58, 71, 195
additivity property of measures 256
additivity property of tree metrics 357
additivity rule 303
adjunct of a mapping 192
algebra of finite type 59
algebra of type $\theta$ 59
algebra type 59
alphabet 26
annulus algorithm 406
antimonotonic mapping 158
approximation space 333
  definable set in an 336
  externally undefinable set in an 336
  internally undefinable set in an 336
  totally undefinable set in an 336
  undefinable set in an 336
Armstrong table 304
Armstrong's rules 301
attribute 46
  domain of an 46
augmentation rule 301
augmenting path for a network 115
axiom of choice 34
axioms for partition entropy 307

Baire space 233
basis for a subset of a dissimilarity space 407
Bell numbers 546
bi-dual collection of a collection 15
bijection 16
binary attribute
  interval 216
  level 216
binomial coefficient 42
Bolzano-Weierstrass property of compact spaces 241
Boolean algebra 192
  morphism 194
Boolean function 196
  binary 203
    $i$-negative 210
    $i$-positive 210
    cover of a 207
    implicant of a 204
    minimal cover of a 207
    prime implicant of a 205

conjunctive normal form of a 202
Boolean function (cont.)
disjunctive normal form of a 199
partially defined 210
simple 196
standard disjunctive coefficients of a 201
standard disjunctive normal form of a 201
Boolean projection function 196
border defined by a closure 170
Borel set 252, 451
Borel-Cantelli lemma 271
boundary of a set 334
bounded set 134
bounded set in a metric space 355
box-counting dimension 479
Buneman's inequality 352

candidate objects 286
capacity of a cut 114
capacity of an edge 111
cardinality of a set 23
carrier of an algebra 59
Cartesian product of two sets 9
centroid of a set 496
closed function 267
closed set 226
closed set of an algebra 63
closed sphere 354
closure of a set of attributes under a set of functional dependencies 303
closure operator 140
closure system on a set $S$ 139
cluster 495
clustering 495
agglomerative hierarchical 495
complete-link 503
dissimilarity conformant to a 517
divisive hierarchical 496
exclusive 495
extrinsic 495
group average method 505
hierarchical 495
intrinsic 495
partitional 495
silhouette of a 520
single-link 500
validity of 522
clustering function 516
$\kappa$-forcing with respect to a 517
$k$-means 516
$k$-median 516
consistent 517
rich 517
scale-invariant 517
cluster point 236
cofinite subset of an infinite set 22
collection of neighborhoods 226
collection of preimages of a set 15
collection of sets 4
decomposition of a 327
dually hereditary 543
hereditary 7, 543
independence number of a 562
independent 562
refinement of a 7
set product of a 25
trace of a 8
witness set of a 326
commutative groups 60
commutative ring 61
complement 5
complemented lattices 187
completeness of Armstrong's axioms 304
components of an ordered pair 7
composition of functions 17
conclusion of a rule 301
conditional attribute 338
confusion matrix of two clusterings 521
congruence of an algebra 62
conjugate of an integral partition 538

connected component of an element 245
consensus of terms 204
containment mapping 16
contingency matrix of two partitions 72
continuous function 241
contraction 445
convex hull of a set 574
convolution product 163
core of a table 296
cost scheme of editing 401
countable set 35
cover of a set 34
crisp set 334
cut 104, 114
cycle in a graph 85
cycle of an element 38

decision attribute 338
decision system 211, 338
  classification generated by a 339
  consistent 339
  decision function of a 339
  deterministic 339
  difference set of a positive and a negative example in a 219
  inconsistent 339
  negative patterns of a 212
  nondeterministic 339
  positive patterns of a 212
  pure 339
definiteness of dissimilarities 352
degree of membership 333
deletion from a sequence 398
dendrogram 373
density constraint 328
density function of a set function 325
density of a collection 561
derangement 544
derived set 235
difference of sets 5
differential constraint 328
digraph 86
  acyclic 87
    ancestor of a vertex in an 87
    descendant of a vertex in an 87
  cycle in a 87
  destination of an edge in a 86
  edge in a 86
  finite 86
  forest as a 87
  in-degree of a vertex in a 86
  length of a path in a 87
  node in a 86
  out-degree of a vertex in a 86
  path in a 87
  path that joins two vertices in a 87
  simple path in a 87
  source of an edge in a 86
  vertex in a 86
dimension of a sequence of ratios 486
dimensionality curse 459
disjoint sets 5
dissimilarity on a set 351
dissimilarity space 352
distance 353
divisibility relation 130
dualization 137
dual statement 137

edge cover 117
edge saturated by a flow 112
editing functions 399
edit transcript 399
  cost of an 401
empty multiset 45
empty topological space 226
endomorphism of an algebra 61
endpoints of an edge 79
equinumerous sets 22
equivalence class 31
essential prime implicant 208
Euclidean norm on $\mathbb{R}^n$ 68
evenness of dissimilarities 352

exact association rules 283
extended dissimilarity 356

factorial 42
Ferrers diagram 538
field 61
field of sets on $S$ 252
filter of a lattice 220
finite algebra 59
finite algebra type 59
finite character collection 8
finite intersection property 239
finite set 22
first axiom of countability for topological spaces 238
fixed point of a function 446
flow 112
Forgy's algorithm 513
format of a matrix 68
four-point inequality 352
Fréchet isometry 454
full hypergraph 118
function 13
  choice 34
  empty 13
  image of an element under a 15
  image of a set under a 20
  indicator 18
  inverse image of a set under a 20
  kernel of a 30
  left inverse of a 17
  pairing 36
  partial 15
  right inverse of a 17
  selective 34
  total 15
function between two sets 15
function continuous in a point 427
functional dependency 298
  proof of a 301
  schema 300
  table that satisfies a 300
  trivial 300

g-measure on a set 321
Galois connection between posets 190
generalization in a partially ordered set 284
Gini index 313
Gini index of a partition 306
graded hierarchy 370
  ultrametric generated by a 371
grading function for a hierarchy 370
graph 79
  $k$-regular 83
  acyclic 85
  adjacency matrix of a 84
  adjacent vertices in a 79
  automorphism 91
  bipartite 83
  bipartite complete 83
  chromatic number of a 122
  clique in a 91
  coloring of a 122
  complete 90
  complete set of vertices in a 91
  connected 88
  connected component of a 88
  cut set of a partition in a 104
  degree of a vertex in a 80
  directed 86
  distance between two vertices in a 85
  edge in a 79
  edge incident to a vertex in a 79
  endpoints of a path in a 85
  finite 79
  forest as a 92
  Hamiltonian path in a 122
  invariant 92
  isomorphism 91
  links of a partition in a 104
  matching of a bipartite 117
  node in a 79
  numbered 108
  numbering of a 108

of a pair of partitions 145
order of a 79
path in a 85
regular 83
separation of a partition in a 104
triangle in a 85
undirected 79
vertex in a 79
weighted 101
greatest element 134
greatest lower bound 135
group 60
group action on a set 75
groupoid 59

Hamming distance 398
Hasse diagram 131
Hausdorff-Besicovitch dimension of a set 484
heading of a tabular variable 47
heap 98
hereditary set 286
hierarchy on a set 368
hitting set of a hypergraph 120
homeomorphic topological spaces 244
homeomorphism 242
hyperedge of a hypergraph 118
hypergraph 118
hyperplane 68
Hölder condition of exponent $\alpha$ 490

ideal of a lattice 220
identity relation of a set 10
immediate descendant of a vertex 95
incidence matrix 84
incidence matrix of a hypergraph 118
inclusion between collections of sets 4
inclusion rule 301
inclusion-exclusion principle 529
index of an element in a set 106
indicator function of a set 18
indiscernibility relation 295
infimum 135
infinite set 22
infix notation for partial orders 130
injection 16
insertion in a sequence 398
integral flow 116
integral partition of a natural number 538
interior of a set 232
interior system 143
intersecting property of a collection 548
intersection 4
associativity of 5
commutativity of 5
idempotency of 5
intersection of two multisets 45
inverse of an element 60
involutive property 323
involutive property of the complement 193
isolated vertex 81
isometric embedding 452
isometric metric spaces 445
isometry 445
isomorphic graphs 91
isomorphic posets 159
isomorphic semilattices 176
isomorphism of Boolean algebras 194
iteration of a function 445
iterative function system 486
attractor of an 487
invariant set for an 486
similarity dimension of an 487

Jordan-Dedekind condition for posets 152

key of a table 298
Kirkhoff's law 112

Kleitman Inequality 543
Kronecker function 163
Kruskal's algorithm 102

large inductive dimension 462
lattice 177
  Boolean 192
  bounded 178
  complement of an element in a 187
  complementary elements in a 187
  complete 188
    isomorphism 189
    morphism 189
  distributive 184
  interval in a 179
  isomorphism 179
  modular 180
  morphism 179
  projection in a 180
  semimodular 182
  sublattice of a 179
least element 134
least upper bound of a set 135
levelwise algorithm 287
Levenshtein distance between sequences 400
linear combination of an arbitrary set 65
linear space 64
  basis of a 65
  inner product on a 66
  linear combination of a finite subset of a 65
  $n$-dimensional 66
  norm on a 67
  real 64
  set spanning a 65
  set that generates a 65
  subspace of a 65
  zero element of a 64
linearly dependent set 65
linearly independent set 65
Lipschitz function 445
logarithmic submodular function 321
logarithmic supramodular function 321
logical implication between functional dependencies 304
lower approximation of a set 334
lower bound 133
lower box-counting dimension 479

mapping 15
marginal totals of a contingency matrix 73
mass distribution principle 485
matrix on a set 68
maximal element 136
maximal flow 112
maximal subdominant ultrametric for a dissimilarity 374
measurable function 254
measurable space 252
measure 256
  completion of a 272
  generalized 321
  outer 258
    Carathéodory 449
    Hausdorff-Besicovitch 483
    Lebesgue 263
    regular 263
measure space 256
medoid 514
Method I for constructing outer measures 261
metric 352
  $\chi^2$ 394
  discrete 353
  Euclidean 382
  Hausdorff 448
  induced by a norm 382
  Minkowski 382
  Ochïai 393
  Steinhaus transform of a 388
  tree 352, 357
metric space 352

$r$-cover of a 482
$r$-cover of a set in a 480
amplitude of a sequence in a 356
covering dimension of a 473
diameter of a 355
diameter of a subset of a 355
embedding of a 452
searching is a 402
external complexity of 409
internal complexity of 409
separate sets in a 355
minimal cut 114
minimal element 136
minimal hitting set of a hypergraph 120
minimal transversal of a hypergraph 120
minimax inequality for real numbers 77
minterms 199
modularity property of measures 257
monochromatic set 533
monoid 59
monotonic mapping 158
monotonicity of measures 256
monotonicity of the Cartesian product 10
monotonicity of clusterings 505
morphism of posets 158
multicollection 46
multiset 44
carrier of a 44
multiplicity of an element of a 44
multiset difference 55
Munroe's Method II 452
Möbius dual inversion theorem 326

negative region of a set 334
net 441
network 111
Newton's binomial formula 43
non-Shannon entropy 311
normalized matching property 54

observation table 211, 215
occurrence of a sequence 28
occurrence of a symbol 28
one-to-one correspondence 16
open function 267
open set 225
open sphere 355
operation 57
$n$-ary 57
arity of an 57
associative 57
binary 57
unit of a 58
zero of a 58
commutative 57
idempotent 57
inverse of an element relative to an 58
multiplicative inverse of an element relative to an 58
unary 57
zero-ary 57
opposite element of an element 58
orbit of an element 75
order of a family of subsets of a set 472
order of a hypergraph 118
ordered pair 7

parameter space 563
partial order 129
discrete 129
extension of a 160
lexicographic 157
strict 129
trace of a 130
transitive reduction of a 132
partially ordered set 129
partition 32
block of a 32
covering of a 144

Gini index of a 306
partition finer than another 32
set saturated by a 33
Shannon entropy of a 306
path that connects two vertices 85
perceptron 563
parameter space of a 563
threshold of a 563
weight vector of a 563
permutation 38
cyclic 39
cyclic decomposition of a 40
descent of a 40
even 41
inversion of a 40
odd 41
pigeonhole principle 536
pivot 402
poset 129
antichain in a 149
Artinian 151
atom in a 135
border of a subset of a 284
chain in a 148
closed interval in a 162
closure operator on a 191
co-atom in a 135
covering relation in a 132
dimension of a finite 169
dual of a 137
finite 129
graded 153
greatest element of a 134
height of a finite 153
height of an element of a 152
incidence algebra of a 163
incomparable elements in a 149
isomorphism 159
least element of a 134
length of a finite 153
level set of a graded 153
locally finite 162
multichain in a 148
Möbius function of a locally finite 166
negative border of a subset of a 284
Noetherian 151
open interval in a 162
order filter in a 168
order ideal in a 168
positive border of a subset of a 284
realizer of a 169
Riemann function of a locally finite 165
standard example 169
upward closed set in a 229
well-founded 151
well-ordered 150
width of a finite 153
positive region of a set 334
precompact set 442
premises of a rule 301
Prim's algorithm 103
principal ideal 221
product of matrices 71
product of metric spaces 413
product of posets 155
product of the topologies 249
product of topological spaces 249
projection 25
projection of a table 48
projection of a tuple 48
projections of the Cartesian product 50
projectivity rule 303
proper ancestor of a vertex 87
proper descendant of a vertex 87
purity of a cluster 522

quasi-ultrametric 354
query 285
query object 402
quotient algebra of an algebra and a congruence 63
quotient set 33

range query 402

rank of an implicant 204
reduct of a table 296
relation 10
  acyclic 131
  antisymmetric 14
  arity of a 26
  asymmetric 14
  binary 26
  collection of images of a set under a 15
  domain of a 11
  dual class relative to a 15
  empty 10
  equivalence 30
    positive set of an 337
    set saturated by an 31
  full 10
  image of an element under a 15
  inverse of a 11
  irreflexive 14
  $n$-ary 26
  one-to-one 13
  onto 13
  polarity generated by a 190
  power of a 12
  preimage of an element under a 15
  range of a 11
  reflexive 14
  symmetric 14
  ternary 26
  tolerance 32
  total 13
  transitive 14
  transitive closure of a 142
  transitive-reflexive closure of a 143
relation on a set 10
relation product 11
relational database 48
replacement 28
residual network of a network relative to a flow 114
ring 60
  addition in a 60
  multiplication in a 60
  right distributivity laws in a 60
  left distributivity laws in a 60
rough set 334

Schröder-Bernstein theorem 221
second axiom of countability for topological spaces 238
selection criterion 496
selective set 34
self-conjugate partition of an integer 547
semi-metric 353
semigroup 59
semilattice 173
  join 176
  meet 176
semilattice morphism 176
separation properties of topological spaces 247
sequence 25
  Cauchy 439
  components of a 25
  concatenation 26
  convergent 435
  contracting 29
  deletion of a symbol from a 398
  divergent to $+\infty$ 436
  divergent to $-\infty$ 436
  expanding 29
  graphic 81
  infinite 27
    ascending 150
    descending 150
  infix of a 27
  insertion of a symbol in a 398
  length of a 25
  occurrence of a 28
  of sets
    contracting 29
    convergent 29
    limit of a 29
    lower limit of a 29
    monotonic 29
    upper limit of a 29

on a set 25
prefix of a 27
product 26
proper infix of a 27
proper prefix of a 27
proper suffix of a 27
Prüfer 108
subsequence of a 27
substitution of a symbol of a 398
suffix of a 27
unimodal 52
sequential cover of a set 263
set of colors 122
set of colors of a set coloring 533
set of finite sequences on a set 26
set of permutations 39
set of polynomials of an algebra 75
set of tuples of a heading 47
set product 25
set shattered by a collection of concepts 551
Shannon entropy of a partition 306
similarity 445
similarity ratio 445
simple cycle 85
simple function on a set 19
simple hypergraph 118
simple path 85
small inductive dimension 462
soundness of Armstrong's rules 301
spanning subgraph 88
specialization in a partially ordered set 284
Sperner family of sets 118
Sperner system 118, 539
Sperner's theorem 540
square matrix 69
standard transposition 39
state of a relational database 48
Stirling numbers of the first kind 546
Stirling numbers of the second kind 538
strict order 129
strictly monotonic mapping 158
subalgebra of an algebra 63
subcollection 4
subcover of an open cover 238
subdistributive inequalities 184
subgraph 88
subgraph induced by a set of vertices 88
subgroup 63
submodular function 321
submodular inequality 180
submodularity of generalized entropy 321
submonoid 63
subset closed under a set of operations 143
substitution 398
sum of matrices 70
sum of square errors 496
sum of two multisets 45
supervised evaluation 520
supramodular function 321
supremum 135
surjection 16
symmetric difference 6
symmetric matrix 69
system of distinct representatives 53

table of a tabular variable 47
tabular variable 47
target of a functional dependency proof 301
Tarski's fixed-point theorem 221
tolerance 31
topological metric space 424
complete 439
large inductive dimension of a 462
separated $r$-set in a 480
small inductive dimension of a 462

zero-dimensional 463
topological property 243
topological space 225
$T_0$ 248
$T_1$ 248
$T_2$ 248
$T_3$ 248
$T_4$ 248
$r$-separation number of a subset of a 480
arcwise connected 268
border of a set in a 233
clopen set in a 231
closed cover in a 238
compact 239
compact set in a 240
connected 245
connected subset of a 245
continuous path in a 267
cover in a 238
dense set in a 230
disconnected 245
Hausdorff 249
locally compact 241
normal 249
open cover in a 238
precompact 442
regular 249
relatively compact set in a 240
separable 230
separated sets in a 266
set that separates two sets in a 466
subspace of a 229
totally disconnected 247
topologically equivalent metrics 425
topology 225
Alexandrov 229
basis of a 238
cofinite 229
discrete 226
finer 229
indiscrete 226
induced by a metric 424
subbasis of a 236
total order 148
totally ordered set 148
training set 337
transitive set 49
transitivity rule 300
transpose of a matrix 69
transposition 39
transversal of a hypergraph 120
tree 92
binary 97
almost complete 98
complete 97
left son of a vertex in an ordered 98
ordered 98
right son of a vertex in an ordered 98
equidistant 363
minimal spanning 102
root of a 95
rooted 95
height of a 95
height of a vertex in a 95
level in a 95
ordered 97
Rymon 106
spanning 95
triangular inequality 352

ultrametric 352
ultrametric inequality 352
ultrametric space 352
unbounded set 134
uncountable set 35
uniformly continuous function 426
union 4
associativity of 5
commutativity of 5
idempotency of 5
union of two multisets 45
unit matrix 70
unitary ring 61
unsupervised evaluation 520

upper approximation of a set 334
upper bound 133
upper box-counting dimension 479
usual topology on $\mathbb{R}$ 226

value of a flow 112
value of a flow across a cut 114
Vapnik-Chervonenkis (VC) class 553
Vapnik-Chervonenkis dimension 551
vector normal to a hyperplane 68
vertex of a hypergraph 118

weight function 390
weight of an edge 101
well-ordering principle 150
word 26

zero flow 112
zero matrix 70